Published March, 2009

A Genome May Reduce Your Carbon Footprint

This somewhat rhetorical title must excite many scientists, particularly those with ongoing research on biomass, feedstock development, and lignocellulosic breakdown/fermention. With the costs of sequencing rapidly decreasing, and with the infrastructure now developed for almost anyone with access to a computer to cheaply store, access, and analyze sequence information, emphasis will increasingly be placed on ways to apply genome data to real world problems such as reducing dependency on fossil fuel. For the efficient production of bioenergy, this may be accomplished through development of improved feedstocks. This article will consider more closely the impact of very cheap sequence data (approximately 1USD per genome) on improvement of switchgrass (*Panicum virgatum* L.), a perennial grass well suited to biomass production.

Technology

If one could cheaply sequence the genome of a single individual perhaps pulled at random from a population or breeding program, how would the information be put to best use? To answer this question one needs to closely consider the biology of switchgrass which is highly outcrossing and polyploid, and the structure of the sequence data itself. The two most important factors (i) whether the technology includes mate-pair information from clone ends, and (ii) whether it produces reads of adequate length, will determine if an accurate genome assembly results. The results would ideally assemble several haploid genomes present in that one individual such that direct comparisons are possible between homoeologous and homologous chromosomes. This is not as straightforward as it seems and clearly requires deeper sequence coverage than in a haploid or inbred individual. Deeper sequence coverage may be cheap, but the production of genetic and physical maps and genomic libraries with large inserts, may all be required for accurate assembly and would add to the cost.

As an example, the genome sequence of the sea squirt *Ciona intestinalis* was determined by shotgun sequencing to an average depth of 8.5x (Dehal et al., 2002). Due to its high levels of polymorphism (1.2%), 15-fold higher than what is found in humans, haplotypes could be reconstructed based on polymorphism detection and mate-pair information. Over larger regions spanning 10–100 kbp, haplotype reconstruction required estimation methods (Jong et al., 2007). Reconstruction was not possible at a whole-genome level and the resulting sequences were a mosaic of disjointed maternal and paternal haplotypes on the order of 40 kbp. These were then coerced together into a draft reference genome sequence that is an oversimplification of reality but very useful nevertheless.

The assembly problem is even more pronounced in outcrossing plants such as poplar (*Populus tremuloides* Michaux). The poplar genome has a lower rate of polymorphism than sea squirt, but is approximately three times its size with more repetitive DNA. Whole-genome sequencing produced an assembly where ~62 scaffold sequences capture over 50% of the genome (N50 = 62) (Tuskan et al., 2006). Deeper sequencing than the average 7.6x coverage would greatly improve haplotype phase determination and will be required for more complete coverage of many plant genomes. However, the greater than 1.2 million SNPs and small Indels which were discovered in poplar means that the need to resequence diverse lines to capture more diversity is reduced. Substantial common nucleotide variation has already been captured.

Published in The Plant Genome 2:5–8. Published 18 Mar. 2009.
doi: 10.3835/plantgenome2009.02.0004let

677 S. Segoe Rd., Madison, WI 53711 USA
An open-access publication

For switchgrass (*Panicum virgatum* L.) (2n=4x=36) the effective genome size is ~1600 Mbp. This is based on a reported DNA content for cv. Alamo of 3.32 pg/cell and represents two heterozygous genomes. It is approximately 10 times larger than the genome of *C. intestinalis* (Hultquist et al., 1996). To sequence these genomes from a single individual to 8x average coverage using 2 × 150 bp of mate pair information with 454 sequencing technology today would require about $1 million USD in sequencing costs assuming $10,000 USD per 454 run. Even with this sequence technology retrotransposon families and highly repetitive regions would probably not be resolved well (Wicker et al., 2006). An extensive physical map, particularly if it provided enough information to resolve homoeologous genomes, would enable assembly and sequencing of a minimum tiling path of large-insert clones and haplotype determination over a much longer region than would otherwise be possible.

With Genome in Hand

Acquisition of the genetic component of natural variation is or will soon become cheap enough that its incorporation through marker-assisted selection into almost all breeding programs is now before us. These breeding programs will be among the greatest beneficiaries of advances in sequencing technology which represents a small paradigm shift that has resulted from the post-genome era. Large yield and productivity gains through molecular breeding are not dependent on understanding the underlying function of genes, regulatory networks, physiological processes, or biochemical pathways, but will only need to understand the predictive value of the genetic or sequence data for selection of one or more breeding goals. With availability of cheap sequencing capacity neither complete sequence assembly or gene annotation is required to apply these techniques. In a species such as switchgrass there exists a great deal of phenotypic variation derived from latitudinal adaptation across its natural range and local adaptation to soil, temperature, and moisture conditions (Casler et al., 2004; Casler et al., 2007). It is still largely undomesticated and thus large gains might be realized through fixation of beneficial alleles in breeding populations. There are likely to be a few genes with large effects that will dramatically impact yields once incorporated into breeding programs. This has occurred during the domestication of all our grain crops, but it may take just a fraction of the time now. Emphasis will be required in several areas to allow this.

Breeding Methodology

I recently learned that reproductive biologists can achieve *in vitro* fertilization with oocytes harvested and matured from fetuses of pregnant cows and are considering using this to increase genetic gain per unit time and cost (Betteridge et al., 1989; Georges and Massey, 1991). It strikes me as a bit Orwellian, but is no different than the common practice of maize breeders to utilize off season nurseries that allow cycling of up to three generations per year. In a cheap sequence world every effort would be made with switchgrass and many other long lived species, and species with long juvenile phases, to decrease generation time, thereby allowing recombination and selection based on marker or sequence data alone. We might achieve this goal with switchgrass if (i) seed dormancy can be overcome through seed treatments such as removal of the seed coat or applied chemicals (Zarnstorff et al., 1994; Sarath et al., 2006) and (ii) through manipulation of photoperiod. With these approaches two to three generations per year seems achievable.

To the maximum extent allowed by time and cost, phenotyping coupled with genotyping or sequencing would establish and update the predictive value of markers or sequenced haplotypic blocks. This could be a coordinated activity to simultaneously improve several locally adapted populations or populations that may exhibit heterosis (Vogel and Mitchell, 2008) and would resemble current practice involving multiple year, multiple location trials in spaced plantings, swards, and possibly progeny evaluation. Each cycle of phenotypic selection would update genetic models and improve them over time. Genomewide selection (Meuwissen et al., 2001; Bernardo and Yu, 2007), marker-assisted recurrent selection (Edwards and Johnson, 1994; Moreau et al., 1998), and other approaches that efficiently use and reuse marker data and rely on genotyping all individuals with a large set of markers would then be applied. These approaches rapidly increase the frequency of beneficial alleles within breeding populations but do not necessarily fix all beneficial QTL. Optimally the methods utilize the contribution of all marker intervals to estimate breeding values and therefore will be effective when there are large numbers of QTL with minor effects. As these techniques are refined to accommodate more sequence information there will be a point at which emphasis will shift to obtaining reliable phenotypic data on more individuals. The training of new workers and development of high-throughput precision phenotyping over multiple environments by multidisiplinary teams or consortia is needed to realize this goal (Hancock, 2005).

Comparative Genomics

Putting sequence information to good use should make use of comparative genomics and this is already an extremely rich area within the Poaceae. One direct result of cheap sequence would be a far more comprehensive survey of genetic diversity that would guide conservation efforts to preserve germplasm diversity and allow reconstruction of past speciation events at a more detailed level. Another result of access to multiple related genomes would be that similarities between closely related species would allow inference of missing data in a

"Jurassic Park" approach. For example if a draft switchgrass genome assembly does not provide a complete assembly as judged by comparison to an inbred genome such as *Sorghum bicolor* L. (Paterson et al., 2009), foxtail millet [*Setaria italica* (L.) Beauv.], or more closely related grass, it will be possible to infer unresolved regions, including retrotransposon family composition and composition of other abundant repetitive elements.

Comparative approaches would be applied to better understand the molecular basis for differences between species that result in higher or lower yields in different environments. For example adaptation of switchgrass to high temperatures and water stress is due in part to C4 photosynthesis. In contrast to maize, sorghum, and miscanthus, C4 photosynthesis in switchgrass is primarily of the NAD^+-me type, although the genus contains examples of all three C4 subgroups. These different C4 subtypes are correlated with metabolic, anatomical, and physiological differences that appear to have profound affects on fitness and yield potential at both high and low temperatures (Naidu and Long, 2004; Ghannoum et al., 2005). Comparative approaches that catalog the functional differences in coding and noncoding regions of the entire photosynthetic apparatus will be extremely informative among C4 subtypes representing $NADP^+$-me, NAD^+-me, and PEP-carboxykinase subtypes within the Panicoideae.

Using gene collinearity and the known syntenic relationships within the Poaceae allows formulation of new hypothesis based on information present in data repositories such as GRIN, Gramene, GrainGenes, and MaizeDB. How these organizations are to handle vast quantities of new sequence information and best serve their stakeholders will require careful planning. Some methods that facilitate access have already been developed to capture public marker and phenotypic data sets in ways that enable reanalysis of the raw data (Casstevens and Buckler, 2004). This process will be increasingly important for comparative purposes especially within groups where hybridization barriers may be overcome and the underlying genes are directly accessible. The potential for interspecific crosses has not been explored within the genus *Panicum* sect. Virgata and sect. Urvilleanum that includes switchgrass, thus the extent to which these barriers exist is not known. Genes and loci with the most significant effects on, for example, ethanol conversion efficiency identified in distant species will be exploited through identifying both natural and induced functional variation within germplasm accessible to a breeder. If no such variation exists, plant transformation represents a viable alternative.

Synergistic Activity

Advances in genome science will rapidly increase our understanding of many biological processes and create opportunities for feedstock development. There will certainly be new approaches that arise through sequencing the diversity present in the microbial world. Metabolic engineering and systems biology in microbes and model plants such as *Arabidopsis*, rice, and Brachypodium have already identified promising candidates for industrial biotechnology (Somleva et al., 2008). Processes like photosynthesis or cell wall restructuring are amenable to engineering but require model systems that allow rapid, informative and controlled experiments to be conducted. Translational approaches that exploit species with short generational times and that are amenable to transformation will be used, but results from these studies will need to be interpreted with great care as has been repeatedly demonstrated when promising results fail to fulfill expectations under carefully structured field trials.

Symbioses and community level studies will likely be greatly affected by advances in sequencing technology. Mycorrhizal associations play major roles in water and nutrient assimilation in prairie ecosystems (Hartnett and Wilson, 1999). Switchgrass is known to be highly dependent on these mycorrhizal fungi (Brejda et al., 1998). Direct interactions with other species including endophytes and pathogens can be studied at the genome level. More complex community and ecosystem level studies could help us to understand the sum total of commensal relationships that exist and why plant species' diversity is correlated with prairie productivity (Picasso et al., 2008). Optimal assemblages of species including soil biota can be identified for diagnostic purposes or for direct manipulation and may help guide sustainable practices. Evaluation of gene expression during the establishment of symbioses and pathogen infection, or in response to allelochemicals or volatile signals can be subject to whole-transcriptome or signature sequencing. The creation of expression atlases for tissues and cell types with the aid of laser capture microdissection can be harnessed to understand gene-family regulation (Nobuta et al.,2007). Sequencing can also identify regulatory loci by identification of expression QT across different genotypes or with ChiP Seq approaches (Gilad et al., 2008; Visel et al., 2009).

Summary

A dollar genome sequence will provide information highways that will cut across several disciplines and will drive the development of next generation biomass feedstocks, bioproducts, and processes for replacing fossil fuels. New feedstocks will produce sustainable high yields with minimal inputs in regions where competition with food is minimized and will provide ancillary environmental benefits associated with carbon sequestration and environmental remediation. Simultaneous technological advances in metabolic engineering and fermentation systems will also come about through availability of cheap sequence data to microbiologists that will result in increased utilization efficiency of lignocellulosic biomass.

The skills that are required to efficiently utilize sequence information in a post-genomics era are not ones that I learned in graduate school. Training of new and subsequent generations of plant biologists will place more emphasis on cross-disciplinary areas such as bioinformatics, statistics, and quantitative genetics as access to sequence information across all taxa will require it. New infrastructure will be required to develop the human capital, produce new computational tools, and foster the partnerships that can leverage sequence information for feedstock development.

Christian M. Tobias
Research Molecular Biologist
Genomics and Gene Discovery Research Unit
USDA-ARS Western Regional Research Center
800 Buchanan Street, Albany CA 94710
email: christian.tobias@ars.usda.gov

References

Bernardo, R., and J. Yu. 2007. Prospects for genomewide selection for quantitative traits in maize. Crop Sci. 47:1082–1090.

Betteridge, K. J., C. Smith, R. B. Stubbings, K.P. Xu, and W.A. King. 1989. Potential genetic improvement of cattle by fertilization of fetal oocytes *in vitro*. J. Reprod. Fertil. Suppl. 38: 87–97.

Brejda, J.J., L.E. Moser, and K.P. Vogel. 1998. Evaluation of switchgrass rhizosphere microflora for enhancing seedling yield and nutrient uptake. Agron. J. 90:753–758.

Casler, M., K. Vogel, C. Taliaferro, and R. Wynia. 2004. Latitudinal adaptation of switchgrass populations. Crop Sci. 44:293–303.

Casler, M., K. Vogel, C. Taliaferro, N. Ehlke, J. Berdahl, E. Brummer, R. Kallenbach, C. West, and R. Mitchell. 2007. Latitudinal and longitudinal adaptation of switchgrass populations. Crop Sci. 47:2249–2260.

Casstevens, T., and E. Buckler. 2004. GDPC: Connecting researchers with multiple integrated data sources. Bioinformatics 20:2839–2840.

Dehal, P., Y. Satou, R.K. Campbell, J. Chapman, B. Degnan, A. De Tomaso, B. Davidson, A. Di Gregorio, M. Gelpke, D.M. Goodstein, et al. 2002. The draft genome of *Ciona intestinalis*: insights into chordate and vertebrate origins. Science 298:2157–2167.

Edwards, M., and L. Johnson. 1994. RFLPs for rapid recurrent selection. Analysis of Marker Data. Joint Plant Breed. Symp. Ser., Am. Soc. Hort. Sci., CSSA. Madison WI.

Georges, M., and J. Massey. 1991. Velogenetics, or the synergistic use of marker assisted selection and germ-line manipulation. Theriogenology 35:151–159.

Ghannoum, O., J.R. Evans, W.S. Chow, T.J. Andrews, J.P. Conroy, and S. von Caemmerer. 2005. Faster Rubisco is the key to superior nitrogen-use efficiency in NADP-malic enzyme relative to NAD-malic enzyme C4 grasses. Plant Physiol. 137:638–650.

Gilad, Y., S.A. Rifkin, and J.K. Pritchard. 2008. Revealing the architecture of gene regulation: The promise of eQTL studies. Trends Genet. 24:408–415.

Hancock, J. 2005. Who will train plant breeders in the U.S. and around the world? Proceedings of the symposium: Plant breeding and the public sector. March 9–11, 2005. Michigan State University.

Hartnett, D.C., and G.W.T. Wilson. 1999. Mycorrhizae influence plant community structure and diversity in tallgrass prairie. Ecology 80:1187–1195.

Hultquist, S., K. Vogel, D. Lee, K. Arumuganathan, and S. Kaeppler. 1996. Chloroplast DNA and nuclear DNA content variations among cultivars of switchgrass, *Panicum virgatum* L. Crop Sci. 36:1049–1052.

Jong, H., M. Waterman, and L. Li. 2007. Diploid genome reconstruction of *Ciona intestinalis* and comparative analysis with *Ciona savignyi*. Genome Res. 17:1101–1110.

Meuwissen, T., B. Hayes, and M. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829.

Moreau, L., A. Charcosset, F. Hospital, and A. Gallais. 1998. Marker-assisted selection efficiency in populations of finite size. Genetics 148:1353–1365.

Naidu, S.L., and S. P. Long. 2004. Potential mechanisms of low-temperature tolerance of C4 photosynthesis in *Miscanthus x giganteus*: an *in vivo* analysis. Planta 220:145–155.

Nobuta, K., R. C. Venu, C. Lu, A. Belo, K. Vemaraju, K. Kulkarni, W. Wang, M. Pillay, P. J. Green, G. Wanget. al.. 2007. An expression atlas of rice mRNAs and small RNAs. Nat. Biotechnol. 25:473–477.

Paterson, A.H., J. E. Bowers, R. Bruggmann, I. Dubchak, J. Grimwood, H. Gundlach, G. Haberer, U. Hellsten, T. Mitros, A. Poliakovet, et al.. 2009. The *Sorghum bicolor* genome and the diversification of grasses. Nature 457:551–556.

Picasso, V.D., E. C. Brummer, M. Liebman, P. M. Dixon, and B. J. Wilsey. 2008. Crop species diversity affects productivity and weed suppression in perennial polycultures under two management strategies. Crop Sci. 48:331–342.

Sarath, G., P. Bethke, R. Jones, L. Baird, G. Hou, and R. Mitchell. 2006. Nitric oxide accelerates seed germination in warm-season grasses. Planta 223:1154–1164.

Somleva, M.N., K. D. Snell, J. J. Beaulieu, O. P. Peoples, B. R. Garrison, and N. A. Patterson. 2008. Production of polyhydroxybutyrate in switchgrass, a value-added co-product in an important lignocellulosic biomass crop. Plant Biotechnol J. 6:663–678.

Tuskan, G.A., S. Difazio, S. Jansson, J. Bohlmann, I. Grigoriev, U. Hellsten, N. Putnam, S. Ralph, S. Rombauts, A. Salamovet. al.. 2006. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science 313:1596–1604.

Visel, A., M. J. Blow, Z. Li, T. Zhang, J. A. Akiyama, A. Holt, I. Plajzer-Frick, M. Shoukry, C. Wright, F. Chenet. al.. 2009. ChIP-seq accurately predicts tissue-specific activity of enhancers. Nature 457:854–858.

Vogel, K., and R. Mitchell. 2008. Heterosis in switchgrass: biomass yield in swards. Crop Sci. 48:2159–2164.

Wicker, T., E. Schlagenhauf, A. Graner, T. J. Close, B. Keller, and N. Stein. 2006. 454 sequencing put to the test using the complex genome of barley. BMC Genomics 7:275.

Zarnstorff, M., R. Keys, and D. Chamblee. 1994. Growth regulator and seed storage effects on switchgrass germination. Agron. J. 86:667–672.