REVIEWS

Genomics of cellulosic biofuels

Edward M. Rubin[1,2]

The development of alternatives to fossil fuels as an energy source is an urgent global priority. Cellulosic biomass has the potential to contribute to meeting the demand for liquid fuel, but land-use requirements and process inefficiencies represent hurdles for large-scale deployment of biomass-to-biofuel technologies. Genomic information gathered from across the biosphere, including potential energy crops and microorganisms able to break down biomass, will be vital for improving the prospects of significant cellulosic biofuel production.

The capture of solar energy through photosynthesis is a process that enables the storage of energy in the form of cell wall polymers (that is, cellulose, hemicellulose and lignin). The energy stored in these polymers can be accessed in a variety of ways, ranging from simple burning to complex bioconversion processes. The high energy content and portability of biologically derived fuels, and their significant compatibility with existing petroleum-based transportation infrastructure, helps to explain their attractiveness as a fuel source. Despite the increasing use of biofuels such as biodiesel and sugar- or starch-based ethanol, evidence suggests that transportation fuels based on lignocellulosic biomass represent the most scalable alternative fuel source[1]. Lignocellulosic biomass in the form of plant materials (for example, grasses, wood and crop residues) offers the possibility of a renewable, geographically distributed and relatively greenhouse-gas-favourable source of sugars that can be converted to ethanol and other liquid fuels. Calculations of the productivity of lignocellulosic feedstocks, in part based on their ability to grow on marginal agricultural land, indicates that they can probably have a large impact on transportation needs without significantly compromising the land needed for food crop production[2].

Lignocellulosic biofuel production involves collection of biomass, deconstruction of cell wall polymers into component sugars (pre-treatment and saccharification), and conversion of the sugars to biofuels (fermentation) (Fig. 1). Partially because of the historically low demand for biologically based transportation fuels, each step in this process is in the early stages of optimization for efficiency and throughput. The crops from which biomass is currently derived have not been domesticated for this particular purpose and the present methods for saccharification and fermentation are inefficient and expensive. However, the recent and pressing desire to develop alternatives to fossil fuels has made the rapid improvement of biofuel production a high priority, in which biologically derived energy ('bioenergy')-relevant genomic insights and resources will have an important role (Table 1).

Biomass

From the perspective of transportation fuels, plants can be viewed as solar energy collectors and thermochemical energy storage systems. It is the storage of energy in a form that can later be accessed via thermochemical or enzymatic conversion that distinguishes biomass from other renewable energy sources. Cellulosic biomass, sometimes referred to as lignocellulosic biomass, is an abundant renewable resource that can be used for the production of alternative transportation fuels[3]. The three main components of lignocellulose are cellulose, hemicellulose and lignin (Fig. 2), with the relative proportions of the three dependent on the material source[4]. Cellulose, the main structural component of plant cell walls, is a long chain of glucose molecules, linked to one another primarily by glycosidic bonds[5]. Hemicellulose, the second most abundant constituent of lignocellulosic biomass, is not a chemically well defined compound but rather a family of polysaccharides, composed of different 5- and 6-carbon monosaccharide units, that links cellulose fibres into microfibrils and cross-links with lignin, creating a complex network of bonds that provide structural strength[5]. Finally lignin, a three-dimensional polymer of phenylpropanoid units, can be considered as the cellular glue providing the plant tissue and the individual fibres with compressive strength and the cell wall with stiffness[6], in addition to providing resistance to insects and pathogens.

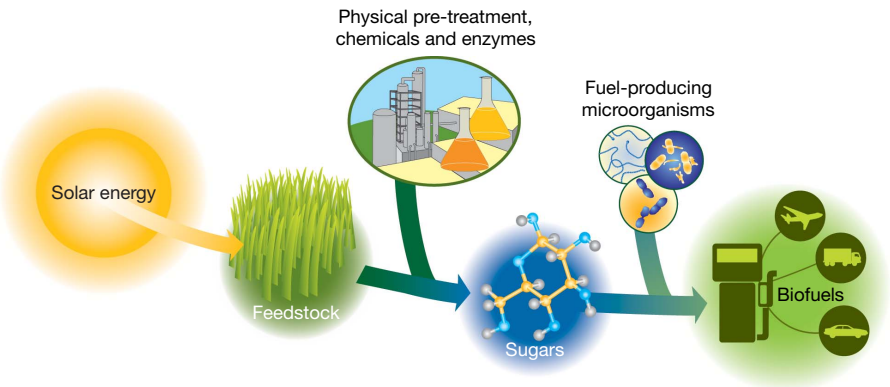

Figure 1 | Biology of bioconversion of solar energy into biofuels. Solar energy is collected by plants via photosynthesis and stored as lignocellulose. Decomposition of the cellulosic material into simple 5- and 6-carbon sugars is achieved by physical and chemical pre-treatment, followed by exposure to enzymes from biomass-degrading organisms. The simple sugars can be subsequently converted into fuels by microorganisms.

[1]DOE Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, California 94598, USA. [2]Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, California 94720, USA.

Table 1 | Bioenergy genomes

Organism	Genome size (megabases)	Status	Reference
Feedstocks and feedstock models			
Populus trichocarpa (poplar)	480	Complete	Ref. 9
Chlamydomonas reinhardtii	120	Complete	Ref. 34
Glycine max (soya bean)	1,200	Draft	–
Manihot esculenta (cassava)	770	In progress	–
Sorghum bicolor	760	In progress	–
Eucalyptus globulus	600	In progress	–
Brachypodium distachyon	355	In progress	–
Zea mays (maize)	2,500	In progress	–
Elaeis guineensis (oil palm)	~3,400	In progress	http://www.checkbiotech.org/green_News_Biofuels.aspx?infoId=15100
Panicum virgatum (switchgrass)	~5,600	In progress	–
Setaria italica (foxtail millet)	~515	In progress	–
Biomass degraders			
Acidothermus cellulolyticus 11B	2.4	Complete	–
Bacillus pumilis SAFR-032	3.7	Complete	Ref. 35
Caldicellulosiruptor saccharolyticus DSM 8903	3.0	Complete	–
Clostridium phytofermentans ISDg	4.8	Complete	–
Clostridium thermocellum ATCC 27405	3.8	Complete	–
Cytophaga hutchinsonii ATCC 33406	4.4	Complete	–
Flavobacterium johnsoniae UW101	6.1	Complete	–
Rubrobacter xylanophilus DSM9941	3.2	Complete	–
Saccharophagus degradans	5.1	Complete	Ref. 36
Thermobifida fusca strain YX	3.6	Complete	Ref. 37
Clostridium cellulolyticum H10	4.0	Draft	–
Elusimicrobium minutum Pei191	1.6	Draft	–
Nectria haematococca/Fusarium solani	51	Draft	–
Phanerochaete chrysosporium	35.1	Draft	–
Postia placenta	33	Draft	–
Sagittula stellata E-37	5.3	Draft	–
Trichoderma reesei/Hypocrea jecorina	33	Draft	–
Cellulomonas flavigena DSM 20109	~4.0	In progress	–
Cellvibrio japonicus Ueda107	~6.0	In progress	–
Fibrobacter succinogenes subsp. *succinogenes* S85	~3.8	In progress	–
Ruminococcus albus	4.0	In progress	–
Teredinibacter turnerae T7902	~2	In progress	–
Termite hindgut community	NA	Complete	Ref. 23
Poplar biomass degrading community	NA	In progress	http://www.jgi.doe.gov/sequencing/lspssseqplans2007.html
Asian longhorned beetle (*Anoplophora glabripennis*) gut community	NA	In progress	http://www.jgi.doe.gov/sequencing/DOEmicrobes2007.html
Bovine rumen community transcriptome	NA	In progress	http://www.energybiosciencesinstitute.org/index.php?option=com_content&task=view&id=159&Itemid=20
Fuel producers			
Clostridium acetobutylicum ATCC 824	4.0	Complete	Ref. 38
Clostridium beijerinckii NCIMB 8052	6.0	Complete	–
Pichia stipitis	15.4	Complete	Ref. 27
Thermoanaerobacter tengcongensis MB4	2.7	Complete	Ref. 39
Zymomonas mobilis subsp. *mobilis* ZM4	2.1	Complete	Ref. 40
Bacillus coagulans 36D1	2.9	Draft	–
Thermoanaerobacter pseudethanolicus 39E	2.4	Draft	–
Clostridium ljungdahlii	~4.0	In progress	–

Bioenergy-relevant organisms for which large-scale genome projects have been completed or are under way are listed. Information on genome projects without references can be found at http://www.ncbi.nlm.nih.gov/sites/entrez?db=genomeprj.

As we can retrospectively view the features that made certain wild plants desirable for domestication thousands of years ago to become today's food crops, we are now prospectively defining criteria to choose plants with potential to serve as dedicated bioenergy crops in the future. These include cell wall composition, growth rate, suitability for growth in different geographical regions, and resource-use efficiencies. With these features in mind, a list of potential bioenergy crops is being developed and targeted for different growing conditions[7]. Most plants assimilate their CO_2 first into a C_3 compound, whereas a smaller subset use a C_4 compound. Plants using C_4 photosynthesis tend to be among the most productive, having higher maximum efficiencies of light, nitrogen and water use in assimilating carbon. The C_4 group of potential energy crops includes various perennial grasses such as switchgrass and *Miscanthus*. These grasses have the advantages of not requiring replanting after a yearly harvest, rapid growth, high biomass density per unit area, and low nutrient and water needs, enabling growth on marginal agricultural land. Disadvantages are that C_4 plants are rare in cold climates and unable to grow at temperatures below 10 °C. In these environments, trees, which exclusively depend on C_3 photosynthesis, provide the only candidate species. The C_3 group of potential energy crops includes trees, such as poplar and eucalyptus, which have relatively rapid growth potential in difficult-to-plough environments. It is highly likely that multiple different energy feedstocks will be deployed depending on latitude, geography, water availability and landowner acceptance.

Until recently, minimal effort has been directed towards optimizing potential energy crops for the generation of transportation fuels. This is in stark contrast to the agronomic development of food crops, which have been domesticated for thousands of years to maximize productivity. Teosinte, the wild precursor to modern maize, was first recognized by Native Americans more than 5,000 years ago as a potential food crop. The domestication of teosinte resulted in its conversion from a wild plant, the characteristics of which had been orchestrated by natural selection maximizing survival and reproduction, into a plant whose morphology and physiology had been extensively altered by artificial selection to increase its nutritional yield and ease of harvest[8]. More recently, selective breeding as well as agronomic advances have resulted in improvement over several orders of magnitude in the nutritional value per acre of modern maize

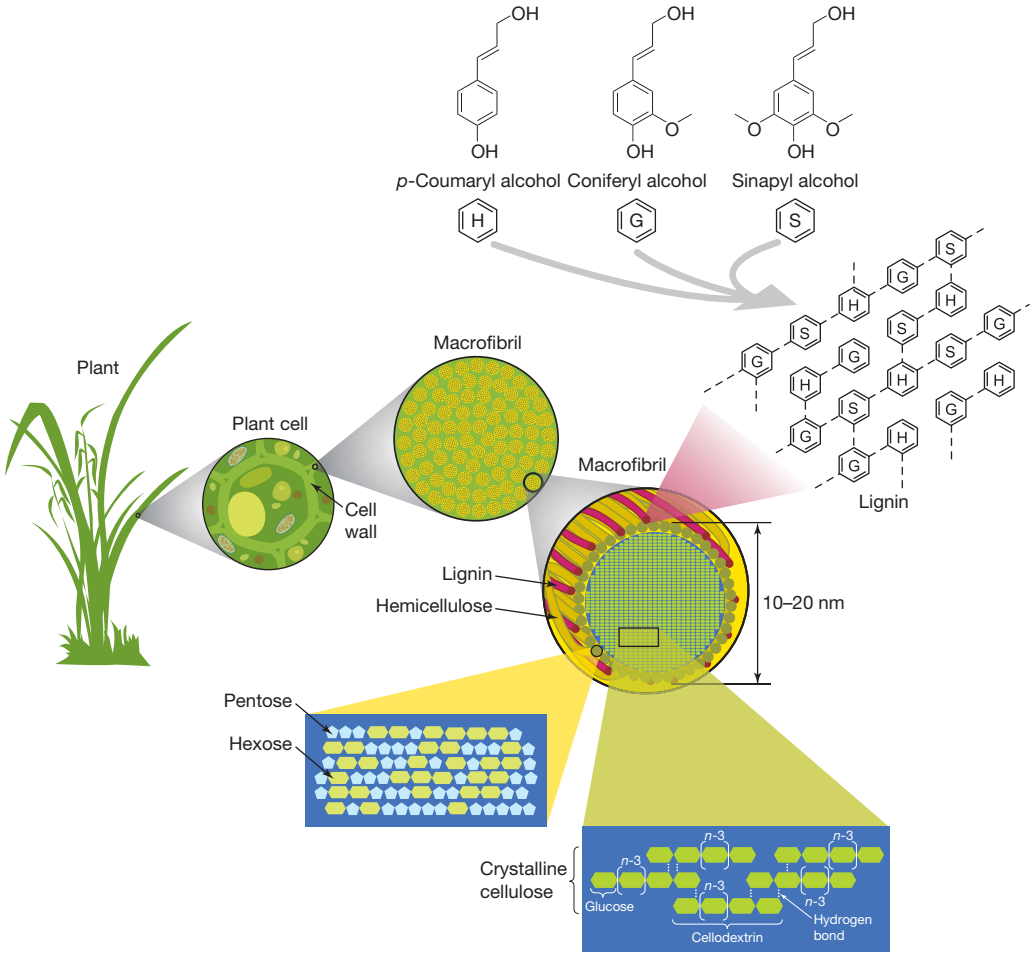

Figure 2 | Structure of lignocellulose. The main component of lignocellulose is cellulose, a β(1–4)-linked chain of glucose molecules. Hydrogen bonds between different layers of the polysaccharides contribute to the resistance of crystalline cellulose to degradation. Hemicellulose, the second most abundant component of lignocellulose, is composed of various 5- and 6-carbon sugars such as arabinose, galactose, glucose, mannose and xylose. Lignin is composed of three major phenolic components, namely *p*-coumaryl alcohol (H), coniferyl alcohol (G) and sinapyl alcohol (S). Lignin is synthesized by polymerization of these components and their ratio within the polymer varies between different plants, wood tissues and cell wall layers. Cellulose, hemicellulose and lignin form structures called microfibrils, which are organized into macrofibrils that mediate structural stability in the plant cell wall.

compared to that of teosinte. Some of the most rapid increases have occurred in the past 40 years, both from advances in agronomic practices and, importantly, from the application of modern genetics. The optimization of bioenergy crops as feedstocks for transportation fuels is in its infancy, but already genomic information and resources are being developed that will be essential for accelerating their domestication. Many of the traits targeted for optimization in potential cellulosic energy crops are those that would improve growth on poor agricultural lands, to minimize competition with food crops over land use.

Populus trichocarpa (poplar), the first tree and potential bioenergy crop to have its genome sequenced (Table 1)[9], illustrates some of the issues and potential of applying genomics to the challenge of optimizing energy crops. The traits for which the genetic underpinnings will be sought in the genomes of bioenergy-relevant plants, such as poplar, include those affecting growth rates, response to competition for light, branching habit, stem thickness and cell wall chemistry. Significant effort will go into maximizing biomass yield per unit land area, because this more than any other factor will minimize the impact on overall land use. One can imagine trees optimized to have short stature to increase light access and enable dense growth, large stem diameter, and reduced branch count to maximize energy density for transport and processing. Trees have evolved with highly rigid and stable cell walls due to heavy selective pressure for long life and an upright habit. Plants domesticated for energy production, with a crop cycle time of only a few years, would have less need for a rigid cell wall than wild plants with lifetimes of a hundred years or more. Alterations in the ratios and structures of the various macromolecules forming the cell wall are a major target in energy crop domestication to facilitate post-harvest deconstruction at the cost of a less rigid plant.

Already, by comparing several of the presently available plant genomes (poplar[9], rice[10,11], *Arabidopsis*[12]; see Table 1) coupled with large-scale plant gene function and expression studies, a number of candidate genes for domestication traits have been identified[13,14]. These include many genes involved in cellulose and hemicellulose synthesis as well as those believed to influence various morphological growth characteristics such as height, branch number and stem thickness[15]. In addition to homology-based strategies, other genome-enabled strategies for identifying domestication candidate genes are being used. These include quantitative trait analysis of natural variation and genome-wide mutagenesis coupled with phenotypic screens for traits such as recalcitrance to sugar release, acid digestibility and general cell wall composition. The availability of high-throughput transgenesis in several plant systems[16] will facilitate functional studies to determine the *in vivo* activities of the large number of domestication candidate genes. Using these strategies, genes affecting features such as plant height, stem elongation and trunk radial growth, drought tolerance, and cell wall stability are but a few of the features that are likely to be identified as targets for domestication

in a fraction of the time required to carry out similar studies unaided by the plant genomes and genomic approaches[17].

Biomass degradation

The breakdown of biomass involves the release of long-chain polysaccharides, specifically cellulose and hemicellulose, and the subsequent hydrolysis of these polysaccharides into their component 5- and 6-carbon chain sugars. Early chemical processes for cellulose degradation depended heavily on acid treatments[5,18], and even today industrial processes for biomass degradation involve heat and acidic conditions and tend to be expensive, slow and relatively inefficient[19]. Furthermore, some of these pre-treatments produce inhibitors (that is, furfural and 5-hydroxymethylfurfural) that decrease the overall yield of the fermentation process.

The human need for efficient breakdown of lignocellulosic biomass for biofuel production is relatively new; however, a variety of organisms have evolved to take advantage of this nutrient source, including the free-living organisms and symbiotic animal–microbe consortia invariably present in biomass-rich environments. Increasing our knowledge of the biochemical machinery used by these organisms for the breakdown of biomass offers new avenues for the development of biologically based processes that could potentially accomplish biomass conversion at an industrial scale. Just as plant genomes are providing us with new approaches for accelerating feedstock domestication, genomics of biomass-degrading microbes and microbial consortia offer a new means to enzyme discovery.

Microbial strategies for degrading lignocellulose are diverse, yet our current understanding of the enzymes involved in these processes is limited to a handful of model organisms such as the fungus *Trichoderma reesei* and the bacterium *Clostridium thermocellum*[20]. Recent genome sequencing of these two biomass degraders, as well as a host of other cellulolytic species (Table 1), has expanded our repertoire of known or potential cellulolytic enzymes. The knowledge of biomass degradation pathways is soon to be increased even more by a number of large-scale genomic surveys under way of isolated organisms and microbial communities known to degrade biomass (Table 1).

Considering the current dependence on acid and heat pre-treatment in the deconstruction of lignocellulose, it is clear that enzymes that are stable and active at low pH values and at high temperatures are of particular value. Thus, enzymes derived from thermophilic and acidophilic organisms known to degrade lignocellulose, such as *Caldicellulosiruptor saccharolyticus* and *Acidothermus cellulolyticus*, hold significant promise for industrial processes[21]. However, the number of sequenced thermoacidophilic organisms is low, partly because of the significant obstacles in cultivating these organisms in the laboratory.

Many novel enzymes and enzyme systems that have evolved to make use of cellulosic biomass are present in difficult-to-culture microbes[22]. This is particularly the case for the communities of microbes inhabiting the guts of lignocellulose-consuming insects and ruminants. Recently, in response to the challenge of studying difficult-to-culture organisms, metagenomic approaches have been developed to access the information encoded in their genomes (Table 1). This typically involves the high-throughput sequencing of DNA extracted directly from the mixtures of organisms present in an environment.

The metagenomic analysis of a wood-feeding higher termite[23] has recently revealed hundreds of genes encoding carbohydrate-active enzymes, many of which were subsequently shown to be biochemically active. From the perspective of an industrial application, the termite hindgut is too complex a model to mimic because it comprises hundreds of microbial species and lignocellulose-degrading enzymes. However, termites tend to feed on complex diets of several plant species, which translate to more complex hindgut communities and enzyme inventories. Thus, examination of this single community reveals enzymes capable of hydrolysing a broad assortment of chemical bonds. The saccharification of a single uniform cellulosic feedstock, particularly one engineered to have a simplified cell wall, will presumably require fewer enzymes and less downstream optimization of gene regulation and enzyme activity in a production host.

The availability of a wide range of naturally occurring lignocellulose-degrading enzymes increases the chances of successful enzyme optimization for industrial processes. Optimization of the saccharification process is crucial because the cost of cellulases remains a key barrier to economical production of biofuels[3]. A more diverse set of candidate enzymes—identified through a combination of conventional cultured microbial genome studies coupled with environmental prospecting methods such as metagenomics—improves the likelihood of obtaining enzymes with activities and stability suited to a variety of industrial processes.

Biofuel production

Although the final steps of cellulosic ethanol production will require much of the same infrastructure developed for the production of sugar- and starch-based ethanol, changes will need to be made to exploit the diversity of sugars generated from the breakdown of biomass. Whereas the conversion of starch-based feedstock results primarily in hexoses, the products from degradation of lignocellulosic biomass, composed in part of hemicellulose, include large amounts of the pentose sugars D-xylose and L-arabinose. In contrast to the hexose sugars, the pentose sugars cannot be fermented by wild-type *Saccharomyces cerevisiae*[24]. Another factor limiting ethanol yield is the toxicity of ethanol to the fermenting host. Most fermenting organisms such as *S. cerevisiae* cannot tolerate ethanol concentrations exceeding 25% (v/v)[25], resulting in a product that must then be concentrated through distillation. Distillation represents an expensive and energy-intensive step in ethanol production.

These limitations in ethanol production have generated significant interest in developing new organisms able to exploit fully the breakdown products of lignocellulosic biomass as well as tolerate the products of fermentation[26]. *Pichia stipitis* represents one yeast species of relevance to biofuel research based on its natural ability to ferment the pentose sugar xylose. Its recently sequenced genome revealed insights into the metabolic pathways responsible for this process[27], and investigators are already working on the optimization of this pathway in *P. stipitis* as well as on the construction of systems for the heterologous expression of *P. stipitis* genes. *Escherichia coli* has already been genetically engineered for the conversion of all hexose and pentose sugars present in hemicellulose polymers[28]. The resulting genetically engineered strain has an ethanol production rate similar to yeast[29], and its ethanol tolerance has been increased by selection on enrichment media[30]. In the future, genomics and pathway engineering should considerably facilitate the development of a variety of organisms able to use the full repertoire of cellulosic and hemicellulosic sugars and tolerate high ethanol concentrations to optimize ethanol yields.

Although ethanol production facilities and distribution centres are rapidly increasing in number, ethanol as a transportation fuel has several disadvantages that encourage the pursuit of better suited alternative biofuels in the future. Disadvantages of ethanol as a biofuel include: (1) high solubility of ethanol in water, which necessitates an energy-intensive distillation step; (2) diminished energy per gallon content of ethanol compared to petroleum; (3) difficulties in the distribution of ethanol via today's pipeline infrastructure owing to its hygroscopic nature; and (4) incompatibilities with many current vehicles at higher blending volumes. Thus, work is under way to enable the synthesis of next-generation biofuel products, such as higher-chain alcohols, alkanes, and other molecules with structures and activities closer to that of petroleum and diesel. Butanol, for example, can be used in conventional engines either in pure form or mixed with petroleum, can be transported by existing pipelines, and has a higher energy content than petroleum[31]. A number of organisms, such as *Clostridium acetobutylicum*, produce butanol

from sugars via the ABE (acetone, butanol, ethanol) fermentation and can be manipulated to produce mostly butanol[31]. *E. coli* has recently been engineered to produce isobutanol and other alcohols via a non-fermentative pathway that may be more readily adapted to large-scale production, using heterologous expression of *Lactococcus lactis* and *Bacillus subtilis* genes[32]. The genomics-enabled construction of partial or fully artificial biological systems through 'synthetic biology' approaches will be key in developing efficient, inexpensive biofuel production systems[33].

Summary

This perspective has focused on biofuels derived from lignocellulosic feedstocks; however, another potential source of transportation fuels is biodiesel, derived from oil-producing plants and microbes. Predominant sources of biodiesel presently include soya bean, rapeseed and palm oils. The genomes of soya bean and oil palm are both currently being sequenced (Table 1), information from which will be called on to increase oil production and facilitate plant growth in a broader range of habitats. Oil-producing algae are also under active study as an alternative source for biodiesel production. One of the major advantages of algae is their growth in liquid, negating the issue of potential competition between food and energy crops for land use. *Chlamydomonas reinhardtii* is one of a series of algal species whose genomes are becoming available for analysis[34]. Considering the magnitude of the transportation fuel problem, biodiesel is likely to contribute at least in part to the solution, and genomic research on oil producers will aid in making biodiesel production cheaper and more efficient.

Placing a man on the Moon and sequencing the human genome represented large-scale technologically challenging projects. In both these instances, substantial research and development efforts resulted in technological advances unforeseen at the projects' initiation. The strategies that were pioneered in sequencing and interpreting the human genome for the improvement of human health are now poised to be an important contributing technology in the challenge to develop environmentally and socially acceptable alternatives to fossil fuels.

1. Hill, J., Nelson, E., Tilman, D., Polasky, S. & Tiffany, D. Environmental, economic, and energetic costs and benefits of biodiesel and ethanol biofuels. *Proc. Natl Acad. Sci. USA* **103,** 11206–11210 (2006).
2. Tilman, D., Hill, J. & Lehman, C. Carbon-negative biofuels from low-input high-diversity grassland biomass. *Science* **314,** 1598–1600 (2006).
   A demonstration of why grassland perennials, such as switchgrass, are superior for biofuel production when compared to crops that presently serve as food crops, such as soya bean or maize.
3. Himmel, M. E. *et al.* Biomass recalcitrance: engineering plants and enzymes for biofuels production. *Science* **315,** 804–807 (2007).
4. Reddy, N. & Yang, Y. Biofibers from agricultural byproducts for industrial applications. *Trends Biotechnol.* **23,** 22–27 (2005).
5. van Wyk, J. P. Biotechnology and the utilization of biowaste as a resource for bioproduct development. *Trends Biotechnol.* **19,** 172–177 (2001).
   A discussion of biowaste as a potential source of lignocellulose for biofuel production.
6. Del Rio, J. C., Marques, G., Rencoret, J., Martinez, A. T. & Gutierrez, A. Occurrence of naturally acetylated lignin units. *J. Agric. Food Chem.* **55,** 5461–5468 (2007).
7. Sanderson, K. US biofuels: a field in ferment. *Nature* **444,** 673–676 (2006).
8. Doebley, J. F., Gaut, B. S. & Smith, B. D. The molecular genetics of crop domestication. *Cell* **127,** 1309–1321 (2006).
9. Tuskan, G. A. *et al.* The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **313,** 1596–1604 (2006).
10. Yu, J. *et al.* A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science* **296,** 79–92 (2002).
11. Goff, S. A. *et al.* A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science* **296,** 92–100 (2002).
12. The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408,** 796–815 (2000).
13. Kalluri, U. C., Difazio, S. P., Brunner, A. M. & Tuskan, G. A. Genome-wide analysis of Aux/IAA and ARF gene families in *Populus trichocarpa*. *BMC Plant Biol.* **7,** 59 (2007).
14. Busov, V. B., Brunner, A. M. & Strauss, S. H. Genes for control of plant stature and form. *New Phytol.* **177,** 589–607 (2008).
15. Ragauskas, A. J. *et al.* The path forward for biofuels and biomaterials. *Science* **311,** 484–489 (2006).
16. Filichkin, S. A. *et al.* Efficiency of gene silencing in *Arabidopsis*: direct inverted repeats vs. transitive RNAi vectors. *Plant Biotechnol. J.* **5,** 615–626 (2007).
17. Dinus, R. J., Payne, P., Sewell, M. M., Chiang, V. L. & Tuskan, G. A. Genetic modification of short rotation poplar wood properties for energy and fiber production. *Crit. Rev. Plant Sci.* **20,** 51–69 (2001).
18. LaForge, F. B. & Hudson, C. S. The preparation of several useful substances from corn cobs. *J. Ind. Eng. Chem.* **10,** 925–927 (1918).
19. Mosier, N. *et al.* Features of promising technologies for pretreatment of lignocellulosic biomass. *Bioresour. Technol.* **96,** 673–686 (2005).
20. Gilbert, H. J. Cellulosomes: microbial nanomachines that display plasticity in quaternary structure. *Mol. Microbiol.* **63,** 1568–1576 (2007).
21. Viikari, L., Alapuranen, M., Puranen, T., Vehmaanpera, J. & Siika-Aho, M. Thermostable enzymes in lignocellulose hydrolysis. *Adv. Biochem. Eng. Biotechnol.* **108,** 121–145 (2007).
22. Hugenholtz, P. Exploring prokaryotic diversity in the genomic era. *Genome Biol.* **3,** REVIEWS0003 (2002).
23. Warnecke, F. *et al.* Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. *Nature* **450,** 560–565 (2007).
    A metagenomic study of an invertebrate gut microbial community involved in lignocellulolytic degradation.
24. van Maris, A. J. *et al.* Alcoholic fermentation of carbon sources in biomass hydrolysates by *Saccharomyces cerevisiae*: current status. *Antonie Van Leeuwenhoek* **90,** 391–418 (2006).
25. Wang, M., Zhao, J., Yang, Z. & Du, Z. Electrochemical insights into the ethanol tolerance of *Saccharomyces cerevisiae*. *Bioelectrochemistry* **71,** 107–112 (2007).
26. Georgieva, T. I., Mikkelsen, M. J. & Ahring, B. K. High ethanol tolerance of the thermophilic anaerobic ethanol producer *Thermoanaerobacter* BG1L1. *Central Eur. J. Biol.* **2,** 364–377 (2007).
27. Jeffries, T. W. *et al.* Genome sequence of the lignocellulose-bioconverting and xylose-fermenting yeast *Pichia stipitis*. *Nature Biotechnol.* **25,** 319–326 (2007).
28. Ohta, K., Beall, D. S., Mejia, J. P., Shanmugam, K. T. & Ingram, L. O. Genetic improvement of *Escherichia coli* for ethanol production: chromosomal integration of *Zymomonas mobilis* genes encoding pyruvate decarboxylase and alcohol dehydrogenase II. *Appl. Environ. Microbiol.* **57,** 893–900 (1991).
    A description of the genetic modification of *E. coli*, yielding a strain capable of fermenting pentose and hexose sugars—which are present in lignocellulose—into ethanol.
29. Jarboe, L. R., Grabar, T. B., Yomano, L. P., Shanmugan, K. T. & Ingram, L. O. Development of ethanologenic bacteria. *Adv. Biochem. Eng. Biotechnol.* **108,** 237–261 (2007).
30. Yomano, L. P., York, S. W. & Ingram, L. O. Isolation and characterization of ethanol-tolerant mutants of *Escherichia coli* KO11 for fuel ethanol production. *J. Ind. Microbiol. Biotechnol.* **20,** 132–138 (1998).
31. Durre, P. Biobutanol: an attractive biofuel. *Biotechnol. J.* **2,** 1525–1534 (2007).
32. Atsumi, S., Hanai, T. & Liao, J. C. Non-fermentative pathways for synthesis of branched-chain higher alcohols as biofuels. *Nature* **451,** 86–89 (2008).
33. Lartigue, C. *et al.* Genome transplantation in bacteria: changing one species to another. *Science* **317,** 632–638 (2007).
34. Merchant, S. S. *et al.* The *Chlamydomonas* genome reveals the evolution of key animal and plant functions. *Science* **318,** 245–250 (2007).
35. Gioia, J. *et al.* Paradoxical DNA repair and peroxide resistance gene conservation in *Bacillus pumilus* SAFR-032. *PLoS ONE* **2,** e928 (2007).
36. Taylor, L. E. *et al.* Complete cellulase system in the marine bacterium *Saccharophagus degradans* strain 2-40T. *J. Bacteriol.* **188,** 3849–3861 (2006).
37. Lykidis, A. *et al.* Genome sequence and analysis of the soil cellulolytic actinomycete *Thermobifida fusca* YX. *J. Bacteriol.* **189,** 2477–2486 (2007).
38. Nolling, J. *et al.* Genome sequence and comparative analysis of the solvent-producing bacterium Clostridium acetobutylicum. *J. Bacteriol.* **183,** 4823–4838 (2001).
39. Bao, Q. *et al.* A complete sequence of the *T. tengcongensis* genome. *Genome Res* **12,** 689–700 (2002).
40. Seo, J. S. *et al.* The genome sequence of the ethanologenic bacterium *Zymomonas mobilis* ZM4. *Nature Biotechnol.* **23,** 63–68 (2005).

Acknowledgements I would like to thank S. Tringe, M. Hess, J. Tuskan, P. Hugenholtz, J. Bristow, B. Simmons, S. Long, J. Fruchart-Najib and H. Blanch for their input to the manuscript. This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Berkeley National Laboratory under contract number DE-AC02-05CH11231, Lawrence Livermore National Laboratory under contract number DE-AC52-07NA27344, and Los Alamos National Laboratory under contract number DE-AC02-06NA25396.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence should be addressed to E.M.R. (emrubin@lbl.gov).