Available online at www.sciencedirect.com
ScienceDirect
Current Opinion in Chemical Biology

Natural products from synthetic biology

Wayne Mitchell[1,2]

DNA sequencing has uncovered microbial secondary metabolic potential that never surfaced in fermentation based screens. Deep and cheap sequencing of a genus such as *Streptomyces* can rapidly expose hundreds of metabolic genes and operons. Meanwhile, synthetic biologists, in their quest to engineer advanced biofuels, are mastering metabolic engineering. Natural products, a reliable source of new therapeutic leads for many years, have fallen into disfavor with drug discoverers partly because these molecules are rarely available as pure compounds and sourcing is often problematic. The convergence of next generation sequencing and synthetic biology, along with less spectacular progress in analytic technologies such as mass spectroscopy, opens the door to the creation of large, reliable libraries of pure natural products for drug discovery.

Addresses
[1] Experimental Therapeutics Centre, Agency for Science and Technology Research (A*STAR), 31 Biopolis Way, #03-01 Nanos, Biopolis, Singapore 138669, Singapore
[2] Nanyang Technological University, School of Computer Engineering, Division of Information Sciences, 50 Nanyang Avenue, Singapore 639798, Singapore

Corresponding author: Mitchell, Wayne (wmitchell@etc.a-star.edu.sg)

Current Opinion in Chemical Biology 2011, **15**:505–515

This review comes from a themed issue on
Next Generation Therapeutics
Edited by Alex Matter and Thomas H. Keller

1367-5931/$ – see front matter

DOI 10.1016/j.cbpa.2011.05.017

Introduction

A starting point for drug discovery revival

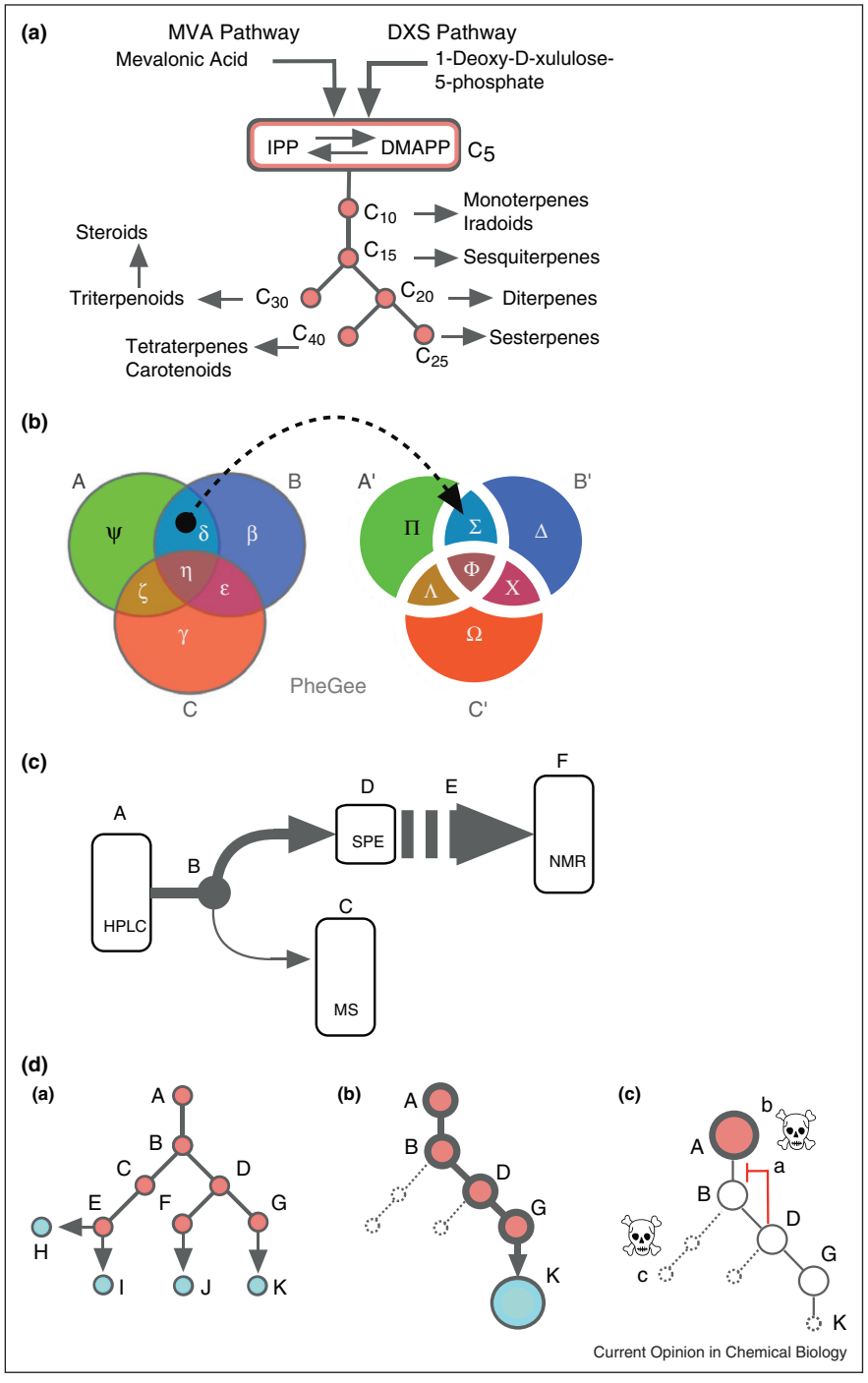
''. . ..there is plenty still to do.''
Sir David Hopwood [1]

Natural products (NPs), derived from the secondary metabolism of bacteria, fungi and plants [2], are chemically distinct from synthetic molecules. NPs have more chiral centers, more oxygen and less nitrogen, sulfur and halogens. The ratio of aromatic rings to total heavy atoms is lower in NPs, the number of solvated H-bond donors and acceptors is higher, and NPs have a broader distribution of mass, log *P* and a greater diversity of ring systems. Surprisingly only 10% of NP structures incur two or more Lipinski Rule of Five violations [3]. Perhaps because of these properties NPs are prominent in the allopathic pharmacopeia, yet the drug industry has moved away from them over the past two decades. Commercial interest in infectious disease indications, an NP stronghold, has flagged. Meanwhile technical innovations such as robotic high throughput screening of single protein targets and combinatorial synthesis have pushed NPs to the margin of the drug discovery process [3].

A renaissance in natural product based drug discovery would be timely. Mounting levels of bacterial drug resistance against existing antibiotics and a concomitant rise in nosocomial infection, especially by the ESKAPE pathogens, portend a genuine public health crisis [4•]. Meanwhile, the drug discovery engine is faltering, as demonstrated by the decline in five year averages for approved new molecular entities (1995–1999: 37.6; 2000–2004: 30; 2005–2009: 20.2) [5]. A 2009 joint study by the European Medicines Agency and the European Centre for Disease Prevention and Control found increasing levels of infection by multi drug resistant bacteria, leading to about 25,000 deaths per annum in the European Union, and, ''a particular lack of new agents . . . against multi-drug-resistant Gram-negative bacteria'' [6]. The Infectious Disease Society of America has set a target of 10 new antibacterial drugs by 2020 to address similar findings in the United States [7]. If this ambitious goal is to be reached, natural products probably will need to play a role. Historically, NPs have been a reliable source of small molecule therapeutics. Of 974 New Chemical Entities, excluding biologics, approved from 1981 to 2006, 39% were NP, NP derived, or fully synthesized NP pharmacophores [8], and natural products continue to deliver. Spirotetrahydro-b-carbolines (spiroindolones) are a class of compounds that have recently produced a novel, potent, orally administered single daily dose anti-malarial candidate. Significantly, the starting point for this discovery was a collection of ~12,000 pure natural products and synthetic compounds with NP-like properties [9•]. Combinatorial libraries that run to over a million compounds are not uncommon, but these collections appear to lack the diversity and enrichment for 'privileged scaffolds' that would typify even a much smaller library of natural compounds [3]. Large collections of pure NPs are rare because they are hard to build by classical fermentation methods, an obstacle that may fall before developments in metabolic engineering.

Synthetic Biology (SB) is the repurposing of living systems for useful ends. SB, philosophically rooted in the engineering paradigm, aims to reduce complex 'natural'

Figure 1



(a) Isoprenoid pathway. Two upstream pathways feed into the pool of the isoprenoid C_5 building blocks, isopentenyl diphosphate (IPP) and dimethylallyl pyrophosphate (DMAPP). Eukaryotes use the Mevalonic Acid pathway to build IPP and DMAPP. Plants have a second route, the D-xululose Pathway. Most prokaryotes only have DSX. Pools of C_{10}, C_{15}, C_{20}, C_{25}, C_{30}, and C_{40} intermediates feed downstream pathways to produce various classes of secondary metabolites. One goal of SB is to build standard chasis optimized for production of NPs originating at each of these nodes.**(b)** Conceptual phenotype–genotype mapping in the PheGee algorithm. The Wenn Diagrams A and A′ illustrate the fact that given a set of genomes, orthologous genes will distribute into sets that are unique to one genome (Ψ, β, λ), are shared between some genomes but not others (δ,ε,ζ), or are ubiquitous (η). Sets of phenotypes distribute similarly ((Π,Δ,Ω),(Σ,X,Λ),(Φ)). Despite confounding effects such as convergent evolution, redundancy, and horizontal gene transfer, the method is surprisingly good at linking molecular phenotypes to candidate genes in prokaryotes.**(c)** A split flow system for natural product structure determination. Extracts from natural or engineered producer strains are separated, for example by HPLC (A). Peaks are captured for analysis, and split into two channels (B), ~5% going to on-line MS analysis (C), the balance diverted solid phase extraction (SPE) media (D), for collection and subsequent off-line analysis by MicroCryoProbe equipped NMR (F). SPE allows analyte to be concentrated by

(i.e. evolved) systems to simplified, reliable, quality-controlled modules, or 'parts', that can be mathematically modeled, manipulated by computer aided design (CAD), 'abstracted' (passed between loosely coupled design and production layers), bolted together to achieve predictable results, and fabricated on an industrial scale [10]. There are compelling reasons to believe SB will strongly influence the bioresearch agenda in the 21st century and in fact may be the vehicle that moves biotechnology into the economic mainstream [11•,12]. Momentum in SB is being driven by so-called 'advanced biofuel' research (cellulosic and algal derived but excluding bioethanol), with very substantial resource inputs from venture capitalists, public equity markets, government subvention and the traditional energy sector totaling well over a billion U.S dollars (USD) [13–16]. Although the focus of the SB biofuel community is the metabolic reengineering of lipid pathways for the production of diesel, jet, automotive fuels and other industrial oils, as a collateral effect the biofuels campaign will enable *generic* synthetic biology technologies including the creation of a technical metabolic engineering knowledge base, the training of a human cohort of practitioners skilled in the art, and the discovery of practical solutions to high level problems in metabolic engineering and their dissemination as general principles.

Artemisinin: an exemplar

The remarkable achievement of the Artemisinin Project provides a compelling example of Synthetic Biology's potential to reinvigorate natural product therapeutics. Artemisinin is a botanical antimalarial isolated from *Artemesia annua*, a wormwood related to *Artemesia absinthium*, and a Traditional Chinese Medicine remedy (*qinqhaosu*, 青蒿素). Like other NPs, artemisinin is biosynthesized in multiple, sequential steps by a suite of functionally related enzymes (a 'pathway') that in bacteria is often recapitulated by a colocalized cluster of corresponding genes (an 'operon'). By transferring plant genes for the artemisinin pathway into a fermentable 'chasis organism' and forcing production of the artemisinin precursors amorpha-4,11-diene and artimisinic acid [17,18], the cost of artemisinin was cut in half, opening access to Artemisinin Combination Therapy (ACT) to low income malaria victims in developing countries [19].

Ingenious general purpose solutions undergird this project's success. Because the isoprenoid artemisinin belongs to a class of compounds that includes many therapeutically interesting members, these advances have broad implications. Isoprenoid biosynthesis branches from a series of precursors that are multimers of C_5 isoprene units [Figure 1a], formed either via the mevalonate dependent MVA pathway (in all eukaryotes) or, via the mevalonate independent DSX pathway (in plants, which have 2 routes to isoprene, and in most prokaryotes). Increasing production of artemisinin precursors, a requirement for increased yield of the downstream product, was complicated by uncharacterized regulatory feedback loops in the circuitry of the native DSX pathway. To overcome this the eukaryotic MVA pathway was introduced into *E. coli* [20], resulting in the hoped for accumulation of the isoprenoid feedstock isopentenyl pyrophosphate (IPP) in engineered cells when grown in mevalonate supplemented medium. However, having bypassed indigenous control, it was then important to re-regulate biosynthesis, since healthy growth and optimal secondary metabolite production result from properly balanced pathway flux. Rational pathway regulation was achieved by codon optimization of imported genes, combinatorial engineering of intergenic regions in target genes [21], by altering enzyme levels through modulation of gene copy number, and by mimicking natural substrate channeling in multizyme complexes by precisely bundling rate limiting enzymes on molecular scaffolds [22••]. Scaffolding is a general solution that has been successfully applied to other systems, including glutaric acid production in an engineered *E. coli* chasis [23].

The artemisinin triumph is not unique. To cite only salient examples, substantial progress has been made in bacterial production of taxol [24••], penicillin has been produced in yeast [25], and a quarter century of inspired work has built up not only a large physical inventory of engineered polyketides and non-ribosomally synthesized peptides, but also a singular understanding of the rules governing their biogenesis [26,27••,28–30].

Blueprint

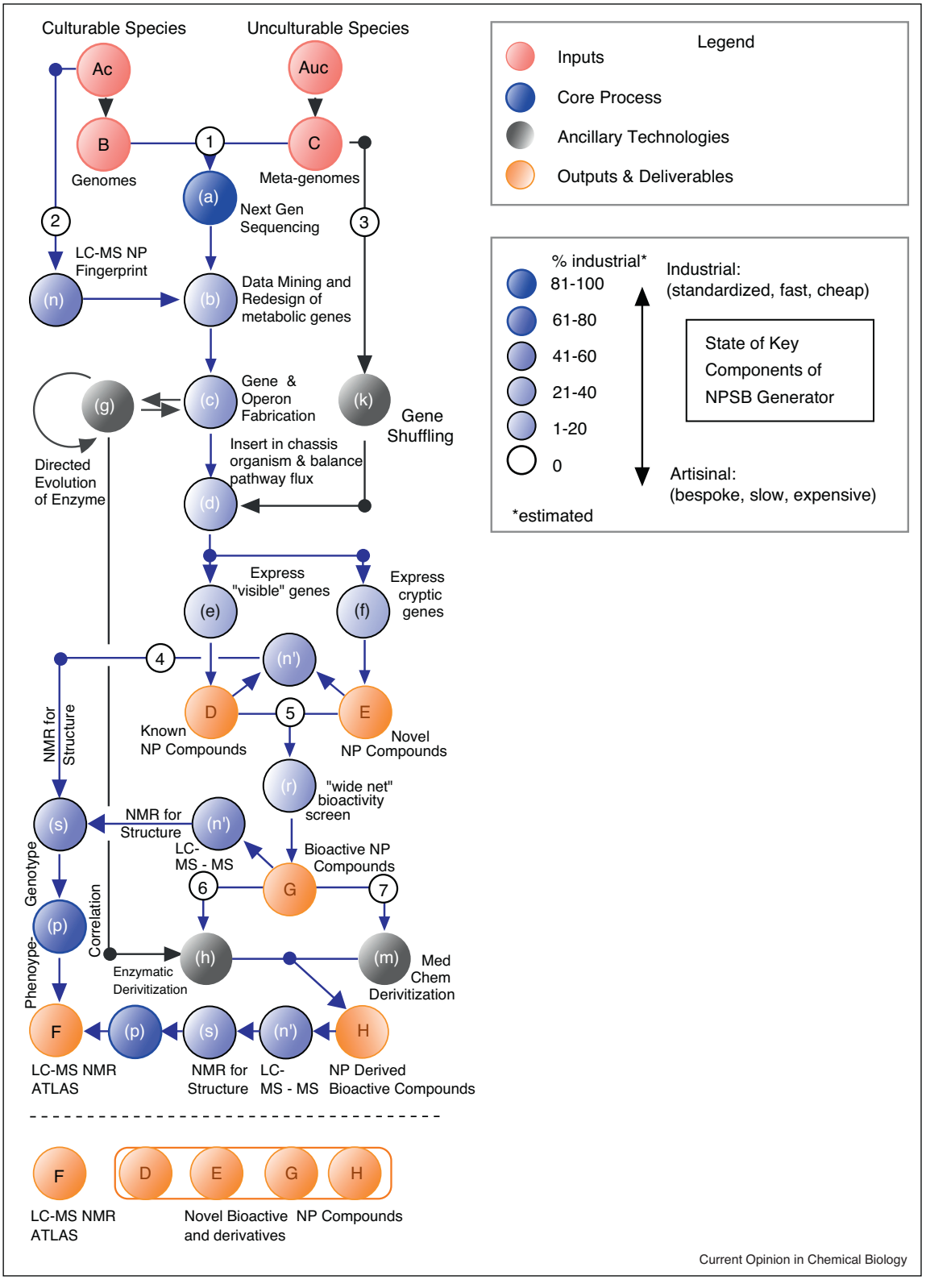
A synthetic biology generator for a natural product 'chemicopia'.

Natural products from synthetic biology

SB promises to do for NP metabolic engineering what automated sequencing and standardized cloning systems have done for molecular biology, triggering a revolution in speed, scale and convenience that will leverage the archive of knowledge accumulated by natural product scientists into the framework of a resurgent NP biology.

(Figure 1 Legend Continued) multiple HPLC runs, thus providing sufficient product for NMR (E). **(d)** Metabolic engineering, the ideal and the reality. (a) An idealized secondary metabolism network forms a tree (a directed acyclic graph), with nodes representing compounds and edges representing reactions, mostly enzyme mediated (intermediates are shown as red circles, end products as blue circles). (b) Conceptually it is straightforward to prune off tree branches leading to undesired products, B–C–E and D–F, thus increasing flux through the path A–B–D–G leading to the desired product, K. (c) In practice this is confounded by many factors. The true graph probably is not acyclic; for example there might be an unexpected negative feedback edge from D to A–B (shown in red at 'a'). High concentrations of D might shut down the pathway upstream and lead to decreased level of the desired product K. Turning off A–B could also cause A to accumulate, and this might be toxic to the host ('b'). Also, the pruning itself might be deleterious. For example if the products of E are essential then deleting E will be fatal ('c').

Figure 2



A process flow diagram for NPSB, Natural Product Compound Library Production using Synthetic Biology. Process inputs are NP producing species (Ac) and their genomic DNA (B), and unculturable NP producers (Auc) and their meta-genomes (C). The backbone process begins with acquisition of whole genome sequence by NGS (path 1, (a)). The secondary metabolome of culturable organisms is sampled by LC–MS fingerprinting to prioritize

How can this momentum be directed most profitably? In terms of drug discovery a concrete goal might be to build a collection of pure NP compounds, reliably and reproducibly sourced in reasonable yield from engineered producer strains, an NP Chemicopia. Such a propitious starting point for drug discovery would be a measurable and useful goal around which resources can be rallied, talent recruited, imaginations engaged and definite plans drawn. A synthetic biology generator to power the engineered biosynthesis of pure NP compound libraries is a realistic goal. Such an engine will combine next generation sequencing, synthetic biology, and small molecule analytics, tightly integrated with complementary informatics and computational design resources. Some components of this engine are fully realized and even commercially available. Others are still in development. But–strikingly–no part of the engine is mere vaporware. It can be built today, albeit with plenty of room for improvement in certain subsystems [Figure 2].

Streptomyces sp.

The genus *Streptomyces* has been the single richest source of NPs beginning with the discovery of actinomycin (*Streptomyces sp*. 1940) and streptomycin (*S. griseus*, 1943) [31], and is an excellent focal point for an initial NPSB campaign. The genus includes over 500 species and thousands of strains and isolates with distinct secondary product capabilities. Thus, the twenty-eight *Streptomyces* genome sequences available [http://www.ncbi.nlm.nih.gov/genomes/MICROBES/microbial_taxtree.html] represent a mere fraction of the genetic and metabolic potential. Between 50% and 90% of secondary metabolism genes identified bioinformatically, amounting to 5% of the very large (~8 megabase) *Streptomyces* genomes, were unknown from previous biochemical surveys [32]. We can conclude that collectively *Streptomyces* probably harbor a considerable genetic potential for secondary metabolism that has been undetected by the classic methods of fermentation and extract screening. Large blocks of genetic dark matter mask the real extent of genetic metabolic potential in the genus.

Next generation sequencing

Next generation sequencing has brought microbial genomics to the lab, and nearly to the benchtop [[33], http://blogs.forbes.com/matthewherper/2011/04/13/ion-torrent-gives-its-dna-sequencing-box-a-boost/] The recent proposal to sequence 1000 bacterial species from prominent NP producing genera is a sound idea [27••]. Arguably, *Steptomyces* alone can consume this effort with good returns. *In lieu* of a large set of fully sequenced and annotated *Streptomyces* genomes, their potential compound space has been estimated from a mathematical model of NP discovery that used a discrete logistic equation modified to include positive and negative feedback (additional screening is encouraged or discouraged by recent success or failure). When fit to an historic time series of NP compound discovery the model's *K*-parameter, corresponding to the total number of discoverable NPs, was estimated at about 150,000 [34•]. One thousand genomes will capture a substantial part of the genes supporting NP metabolism in the genus only if the total potential compound discovery space is much smaller than this and the genomes are non-redundant at these loci (neither likely to be true). A hypothetical 1000 *Streptomyces* Genomes Project therefore should be seen as a preliminary reconnaissance, although one that certainly will expose ample secondary metabolic gene clusters for a robust test of the NPSB engine. Looking beyond the first thousand, sequencing of additional genomes should continue until a plot of novel metabolic gene discovery as a function of the number of genomes sequenced approaches zero slope (or some minimum rate to be decided). Sequencing is cheap. It makes sense to continue sequencing out to the edge of *Streptomyces sp.* diversity. With nearly 100,000 strains in some collections, the project initially can rely on culturable strains and species for material. In the long run, NPSB will want to tap into the much greater genomic potential of the unculturable microbes that make up at least 99% of diversity. This can be accomplished by 'metagenomic' sequencing of environmental DNA. As an example of what can be accomplished, a functional screen of a 4.7 gigabase metagenomic library uncovered three genes that improved the growth of *E. coli* 5.7–6.9 fold in the presence of inhibitors produced in biomass conversion [35]. This is a valuable finding, but the discovery rate is low, an observation that underscores the advantage of the more efficient NPSB scenario where only rationally selected genes and operons will be fabricated.

(Figure 2 Legend Continued) genomes (path2, (n)). Transcriptional data can also be gathered on culturable organisms under different induction conditions (not shown). Next, genomic content is mined to extract metabolic operons and genes; at this step LC–MS fingerprints can be correlated roughly with genomic content. Genes of interest are then redesigned and optimized for expression using the evolving set of synthetic biology design principles (b). The reconfigured operons and genes are sent for physical fabrication (c). Genes for metabolic enzymes can be shunted to a parallel process to extend their synthetic capabilities by directed enzyme evolution (g) and either returned to the backbone process or, employed as derivatizing agents (h). Shuffled metagenomic DNA (path 3, k) can be piped into the process at this point. Physical genes and operons are 'bolted' to an appropriate chassis organism and pathway flux is tuned to optimize production of secondary metabolites (d). In some percentage of cases operons direct the biosynthesis of known or previously unknown metabolites (D and E) that are purified if present at a threshold concentration empirically determined on the basis of convenience, or simply processed as extracts The LC–MS Fingerprinting module is employed again on these compounds or extracts (n′), ideally in conjunction with NMR (s), to create much tighter phenotype–genotype correlations that are systematically accumulated in a database (path 4, (p) and F). The D/E compound set can be filtered by a 'wide net' bioactivity screen such as live cell phenotype screening, and the presumably smaller set of bioactive compounds (path 5, G – not discussed in this review) then purified, and further expanded by medicinal chemistry (path 7, m) or through the action of evolved enzymes (path 6, h) to produce a more diverse set of compounds (H) that will be characterized by LC–MS-SPE-NMR.

Genome mining

The purpose of genome mining is to identify genes responsible for secondary metabolism for subsequent metabolic engineering. Microbial genomics can expose the genetics underlying the biosynthetic potential hidden in cryptic and unexpressed operons ('cryptons'). However, the function of a large percentage of genes, 20–40% in a typical bacterial genome, will be of unknown function, so metabolically interesting cryptons must be culled by bioinformatic analysis. Where protein homology and motif analysis do not suffice, a phenotype to genotype algorithm can be useful. With *n* genomes sequences in hand the cooccurrence of gene homologs define subsets of genomes that range in size from a single member (for a gene found in a single genome) to all *n* members (for a gene present in all genomes). A straightforward mapping between shared homolog subsets and experimentally defined phenotype positive subsets link observed traits to a small number of candidate genes that can then be tested experimentally [Figure 1b]. This simple approach is surprisingly good at identifying gene candidates for molecular traits in bacteria [36]. The transcriptome can be used as a 'genotype' to narrow the list of genes potentially responsible for the strain's secondary metabolic profile (the phenotype), providing both data sets are generated under the same conditions. NGS is emerging as the method of choice for measuring levels of cDNA (mRNA) in prokaryotes, despite the absence of 3′ poly adenylation on their transcripts [37]. *In extremis* very sensitive and sophisticated Hidden Markov Model based tools for the statistical prediction of secondary metabolism genes are also available [38].

Overall, data mining is effective as shown by the rapid exposure of two cryptic sesquiterpene synthases in the recently sequenced *S. clauligerus* genome [39].

Fingerprints and Atlases

The key analytical technologies of this platform, chromatography (usually HPLC), mass spectroscopy (MS), and nuclear magnetic resonance (NMR), have been challenged to meet the demands of accurate NP characterization of bacterial extracts in non-targeted metabolomic screens. Although these technologies are not improving as rapidly as the omic equipment, there is progress. MS for example, while laborious, is obviously many times faster and cheaper than it was just a few years ago. As more genomic sequence is collected, along with corresponding low resolution MS finger prints of bulk metabolites, it may become possible to correlate the two, at least at the level of prioritizing species for deeper scrutiny. MS has also been used repeatedly to generate molecular fingerprints for phenotype–genotype mapping [40••] even in investigations more challenging than NPSB (where a small number of known genes are being investigated in a defined host). Informatics plays a key role. If we assume that initially most NPs will be novel, then a large NMR effort will be required for structure elucidation. As compounds are purified, their MS–MS profile can be determined and stored in an MS–MS profile database (or 'Atlas'). In future rounds of experimentation, routine MS–MS analysis will catch progressively more known structures, taking some of the burden off NMR [41]. MS can also be combined with genomic information and isotopic labeling ('genomisotopics') to link metabolites back to orphan operons as part of the genome mining process [42].

NMR, the lowest throughput component of the NPSB engine, is also the process that demands collection of the most material, leading to longer experimental cycles, and higher cost per compound identified. A technical NMR advance that reduces the sample requirement is the microcryoprobe. Using a 600 MHz instrument and a 1.7 mm microcryoprobe it was possible to identify bacterial derived trisoxazole macrolides extracted from a single host sponge [43]. Methods are evolving that should allow tighter integration of MS and NMR and more automation of the coupled process. A general approach is to split the LC eluent stream when a peak is detected, sending a small percentage to MS and the bulk to Solid Phase Extraction (SPE) media, where compounds are trapped, and potentially concentrated by repeated LC runs, for offline NMR processing [Figure 1c] [44,45]. Not every advance need entail a new machine. Systematic investigation of NP extraction conditions revealed that addition of ammonium fluoride improved ESI-MS performance, with a 2-fold increase in detectable features and 3.5 fold increase in hits in compounds databases [46].

Fabrication

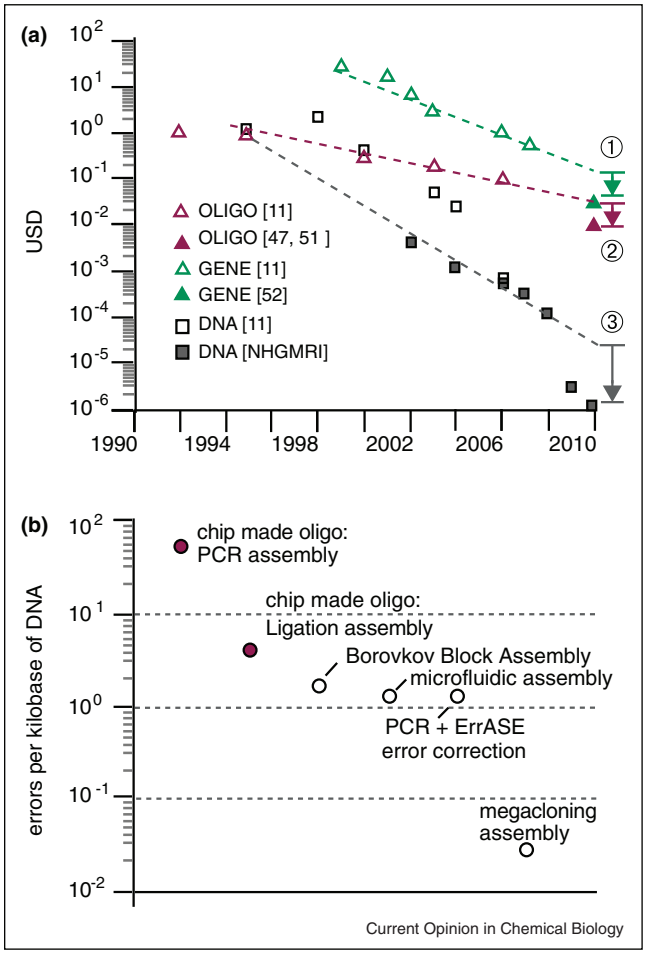
The ability to design and routinely manufacture cheap and reliable DNA components, at least up to the size of bacterial operons, is a fundamental premise of synthetic biology and an absolute requirement of NPSB. Significant progress has been reported. Fabrication has two components: the chemical synthesis of DNA oligo building blocks approximately one hundred bases long and their subsequent assembly into long multi-gene sequences thousands and tens of thousands bases long. The obvious advantage compared to cloning from genomic DNA or cDNA is the total control exerted over the design and sequence of the product, features such as restriction sites and promoters can be inserted or deleted almost at will. More importantly, a fabricated construct can be codon optimized for good expression in a heterologous host, for example in an *E. coli* chasis, an issue especially pertinent to *Streptomyces* where GC content is high. Although the cost of DNA synthesis has not yet followed DNA sequencing down to the vanishing point, recent developments indicate a corner has been turned, and economical gene synthesis now looks feasible [Figure 3a].

The classic method for oligo nucleotide synthesis, 3′ to 5′ addition of nucleoside phosphoramidites tethered to a silica bead substrate, is too expensive for operon scale synthesis. Automated platforms such as the 'MerMade', brought DNA synthesis by this method down to about 0.50 USD per base by 2006, approximately the cost of the underlying chemical reagents. That price is still prohibitive. An ultra miniaturized microfluidic oligo synthesis platform has been reported, which will minimize reagent usage and further lower price, but probably not as low as array synthesized oligos [47] (see below). Cost aside, there are technical limitations in the length and purity of resin-based oligo synthesis that limit their usefulness for SB. Failure at any synthetic step of an individual oligonucleotide truncates further downstream addition lowering total yield for full length oligonucleotides, necessitating labor intensive purification by HPLC, or gel electrophoresis to reliably acquire the desired product.

Assembly of oligos into larger, functional genes and operons is even more expensive. Assembly follows one of two basic strategies [Figure 4a,b]. One method is to design oligo building blocks with unique overlapping ends that will hybridize to form a skeletal, partially double stranded template that then can be filled in by Polymerase Chain Reaction (PCR) to complete the construct. Alternately, a complete spanning set of oligos can be made, then hybridized and stitched together by a ligase enzyme. The second route requires a larger investment in oligos, but has the advantage of bypassing the error prone second-strand fill-in by PCR. *In vitro* constructs up to 500,000 base pairs have been created using refinements of this method [48]. These very large constructs can also be assembled and maintained in yeast, and efficient shuttle tools have appeared, including a three-way yeast/*E. coli*/'other' system [49].

Oligos synthesized on chips by photoactivation or ink jet are much less expensive than resin synthesized polymers, as little as 0.01 USD per base. However, chip synthesis is also inherently more error prone, and, the product oligos are present individually in low concentrations and therefore must be amplified by PCR thus introducing additional sequence defects. Furthermore, pools of chip synthesized oligos are highly complex mixtures promoting inappropriate cross hybridization and misassembly. These problems have been overcome by meticulous computational design of flanking sequences that allowed a total pool of 55,000 unique oligos to be 'amplification sorted' into 10 subpools, each potentially encoding about 100 different genes. The sub pools were individually amplified by sequential, differential PCR, then assembled into complete genes at an final cost of less than 0.10 USD per base. This is a significant, order of magnitude economy [50••,51].

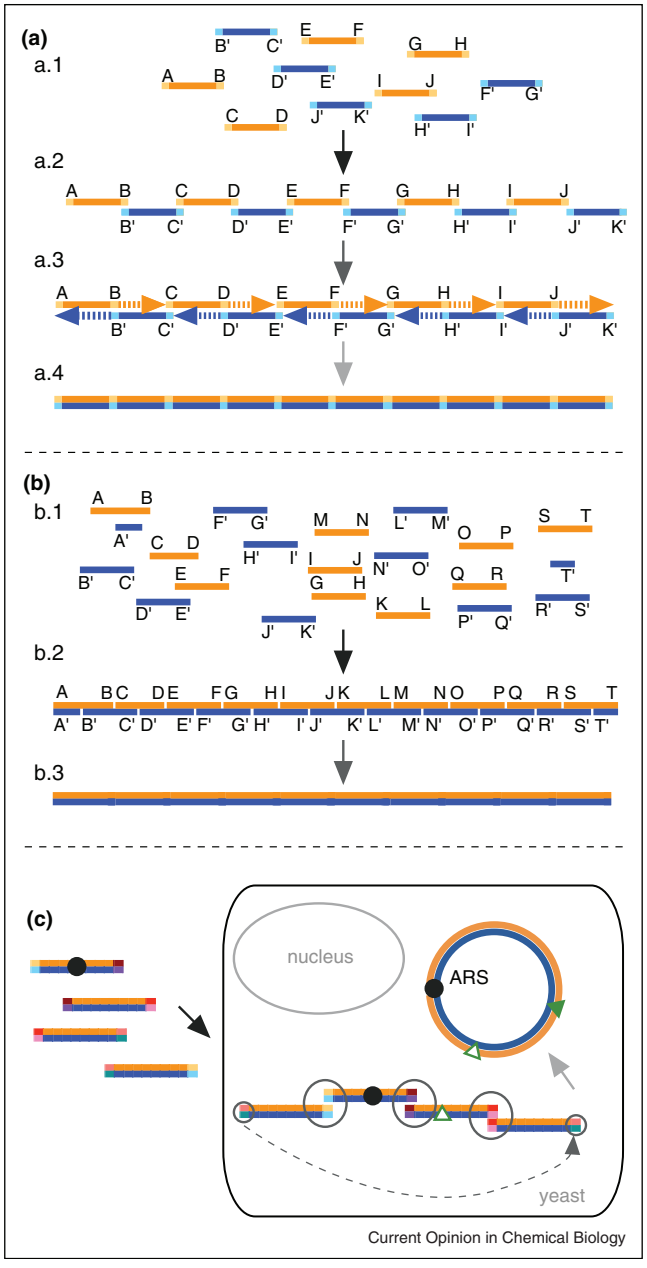
Figure 3



(a) The cost of sequencing, oligo manufacture and gene assembly is declining. Next Generation Sequencing (NGS), shown in open black squares [11•] or closed black squares [Wetterstrand KA. DNA Sequencing Costs: Data from the NHGRI Large-Scale Genome Sequencing Program, www.genome.gov/sequencingcosts. Accessed 06042011] has pushed the cost DNA sequence down to negligible levels. Oligo manufacturing costs, shown in open red triangles [11•] or closed red triangles [47,51] only declined slowly through about 2005 but have dropped an order of magnitude in the past two years. The cost of assembled genes, shown in open green triangles [11•] or closed green triangles [52••] also dropped by an order of magnitude over the same period. These recent declines are marked by downward pointing arrows drawn from the projected trend lines (based on about a decade of data) to the 2010 data points (1,2,3).
(b) Error rates in assembled gene constructs are shown on a per kilobase basis for various error correction technologies. Error rates for uncorrected assemblies are shown as solid purple circles: chip synthesized oligos with PCR assembly and ligation assembly. Error rates for various strategies are shown as open circles: block assembly [52••], microfluidic assembly [47], PCR assembly with ErrASE [54•], and megacloning [53••].

Error propagation is a serious problem for any gene assembly strategy because the final product must direct accurate transcription/translation of a viable protein. Deletions or insertions are likely to cause frame shifts,

Figure 4



Three strategies for assembling genes, operons and genomes from oligonucleotides. **(a)** (A.a.1) Polymerase Assembly. A series of oligos are designed relatively short, overlapping complimentary regions, adhering to a narrow Tm range. Complimentarity is indicated in the figure by a capital letter-letter prime pair, for example oligos A–B and B′–C′. (A.a.2) A linear network of oligos is formed by hybridization, for example A_B–B′_C′–C_D . . . (A.a.3). The single stranded gap regions are filled-in by a polymerase. **(b)** Ligase Assembly. (B.b.1) A spanning set of oligos is designed represent complete coverage of both strands. (B.b.2) The oligos are hybridized to form a double-stranded construct with unjoined 'nicks' at the oligo ends. (B.b.3) The construct is stitched together with ligase enzyme. **(c)** Whole Genome Assembly. Very large constructs (>500,000 base pairs), containing genetic modifications engineered into the constituent oligonucleotides (open green triangle) can be assembled in yeast by homologous recombination (open grey circles and dotted arrow); by including a yeast ARS (solid black circle). These constructs can be maintained as Yeast Artificial Chromosomes, and further modified using the yeast molecular biology tools kit (solid green triangle).

and even frame-preserving non-synonymous substitutions can render the construct useless if the altered amino acid disrupts protein function. The JVCI genome transplantation project was derailed for several months by a single base deletion in the essential replication protein *dnaA* that dysfunctionalized an entire 500,000 base pair whole genome construct [52••]. The workaround has been to painstakingly purify the starting oligos, to sequence the final construct extensively to identify errors, then to use PCR repair, and finally to functionally screen the constructs for activity. These techniques are labor intensive, require skilled practitioners, and drive up the price of finished genes and operons (nearly 1.00 USD per base in 2007 [Figure 3a]). Errors can be *suppressed* pre-assembly by using the highest possibly quality oligos, or errors can be *removed* post-assembly. In a gene assembly strategy called 'megacloning' next generation sequencing (NGS), reimagined, is used as a preparative technique to purify perfect oligos and maximize error suppression [53••]. For post assembly an efficient error correction system called ErrASE has been developed and will be distributed commercially by Life Technologies [54•]. Overall, progress in driving down error rates in assembled genes is impressive [Figure 3b] meaning that far fewer constructs need to be sequenced and functionally tested. As fabrication is a fundamental SB procedure, these advances are crucial.

Metabolic engineering

The importance of balancing the activities of multiple enzymes along a pathway has already been emphasized [Figure 1d]. This also can be accomplished at the level of translation. Riboswitches are natural RNA aptamers that modulate gene expression or translation in response to ligand binding. Often the ligand is a product of the gene circuit being regulated, thus forming a feedback loop. The potential of riboswich control is apparent in the successful reengineering of the add-A mRNA element that operates by closing or opening access to the ribosome binding site in response to binding by adenosine. An orthogonal add-A* was created that was unresponsive to adenine but highly responsive to ammeline and azacytosine, ligands inactive on native add-A [55]. A completely different approach to regulatory engineering has been taken in which parallel evolution of multiple ribosomal binding sites is orchestrated by the addition of mutagenic oligos complimentary to the lagging strand of DNA synthesis that are introduced through cycles of growth and electroporation [56•]. Another approach to translation tuning has been demonstrated in yeast where series of hairpin RNA substrates have been designed that have a

spectrum of susceptibility to the yeast RNase III cleavage enzyme Rnt1p. This spectrum of activity correlated with protein levels when the hairpins where inserted in the 3′ untranslated region (UTR) of a test gene, indicating that the system shows promise as an 'off the self' kit to tune protein expression [57].

Conclusion

Ungent. Urgent. Opportune.

'Snakeoil' has come into English as a derogatory shorthand for 19th century patent medicines that were at best useless, and often frankly harmful. But snake oil or *sheyou* (蛇油) is in fact a standard remedy still in use in Chinese Tradition Medicine (TCM), a *bone fide* natural product, albeit a non-standardized, non-crystalline, molecularly uncharacterized admixture. Just the features that made NPs unattractive to drug developers in the first place. Oil of a different sort is fueling explosive growth in synthetic biology. I have argued here that an SB engine can be built and used to assemble libraries of purified NP compounds that will offer excellent starting points for a revival in drug discovery. Whether *sheyou* itself is or is not efficacious, the persistence of TCM, Ayurveda and other traditional pharmacopeiae [58–60] remind us that NPs represent a rich therapeutic lode waiting to be mined. The creation of pure natural product and natural product derived collections, a natural product Chemicopia, is urgently needed, is feasible, and is opportune.

To move forward the community should embrace a definite and immediate program linked to concrete objectives. A manifesto for the first phase of such a campaign might have the following platform:

1. Sequence *Streptomyces* to the edge of new secondary metabolic gene discovery, 1000 genomes to start.
2. Use synthetic biology to build a collection of pure *Streptomyces* derived natural products, 10,000 to start.
3. Collaborate with medicinal chemists to derivatize these compounds where appropriate to make additional pure semi-natural products, 5000 to start.
4. Make these compounds widely available for screening in drug discovery programs.

Conflict of interest statement

The author is unaware of any conflicts of interest that would materially influence the content of this work.

Acknowledgements

The author thanks the many NP and SB scientists who are not cited in this review owing to the abundance of good work and the dearth of space. WM is supported by a block grant to the Experimental Therapeutics Centre from the Biomedical Research Council of the Agency for Science and Technology Research (A*STAR), Ministry of Trade and Industry, Singapore.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Hopwood DA: *Streptomyces in Nature and Medicine. The Antibiotic Makers*. Oxford; 2007.

2. Dewick PM: *Medicinal Natural Products A Biosynthetic Approach*. Chichester, UK: John Wiley & Sons; 2009.

3. Koehn FE, Carter GT: **The evolving role of natural products in drug discovery**. *Nat Rev Drug Discov* 2005, **4**:206-220.

4. • Boucher HW, Talbot GH, Bradley JS, Edwards JE, Gilbert D, Rice LB, Scheld M, Spellberg B, Bartlett J: **Bad bugs, no drugs: no ESKAPE! An update from the Infectious Diseases Society of America**. *Clin Infect Dis* 2009, **48**:1-12.
This update from clinical scientists documents the deteriorating bacterial multi drug resistance problem in the United States, with particular attention to the ESKAPE pathogens, *Enterococcus faecium, Staphylococcus aureus, Klebsiella species, Acinetobacter baumannii, Pseudomonas aeruginosa, and Enterobacter* species.

5. Kaitin KI, DiMAsi JA: **Pharmaceutical innovation in the 21st century new drug approvals in the first decade, 2000–2009**. *Clin Pharmacol Ther* 2011, **89**:183-187.

6. European Center For Disease Prevention and Control and the European Medicines Agency: *Joint Technical Report. The bacterial challenge: time to react*. EMEA doc. Ref EMEA/576176/2009.

7. The Infectious Disease Society of America: **The 10 X '20 Initiative: Pursuing a Global Initiative to Develop 10 New Antibacterial Drugs by 2020**. *Clin Infect Dis* 2010, **50**:1081–1083.

8. Newman DJ, Cragg GM: **Natural products as sources of new drugs over the last 25 years**. *J Nat Prod* 2007, **70**:461-477.

9. • Rottmann M, McNamara C, Yeung BKS, Lee MCS, Zou B, Russell B, Seitz P, Plouffe D, Dharia NV, Tan J *et al.*: **Spiroindolones, a potent compound class for the treatment of malaria**. *Science* 2010, **329**:1175-1180.
A 12,000 NP and NP-like compound library supplied initial hits leading to a successful medicinal chemistry program and a novel anti-malarial clinical candidate.

10. The Royal Academy of Engineering: *Synthetic Biology: Scope Application and Implications*. London; 2009.

11. • Newcomb J, Carlson R, Aldrich SC: *Genome Synthesis and Design Futures: Implications For The U.S. Economy*. Bio Economic Research Associates for U.S. Department of Energy. Cambridge; 2007.
One of the best and most comprehensive analysis of technological development and potential economic impact of SB. Lucid and informative.

12. Carlson R: **The market value of GM products**. *Nat Biotechnol* 2009, **27**:984.

13. Mouawad J: ***Exxon to invest millions to make fuel from algae***. *New York Times*; July 13, 2009.

14. Gupta P, Baldwin C: **Codexis seen good sign for green IPOs**. *Rueters*; April 23, 2010.

15. Avalos G: **Amyris IPO fetches $85 million**. *Oakland Tribune*; September 28, 2010.

16. Kuo I: **Khosla-backed Gevo raises $107 million in biofuels IPO**. *Reuters*; February 9, 2011.

17. Tsuruta H, Paddon CJ, Eng D, Lenihan JR, Horning T, Anthony L, Regentin R, Keasling JD, Renniger NS, Newman JD: **High-level production of amorpha-4,11-diene, a precursor of the antimalarial agent artemisinin, in Escherichia coli**. *PloS ONE* 2009:1-12 doi: 10.1371/journal.pone.0004489.

18. Anthony JR, Anthony LC, Nowroozi F, Kwon G, Newman JD, Keasling D: **Optimization of the mevalonate-based isoprenoid biosynthetic pathway in Eschericia coli for production of the anti-malarial drug precursor amorpha-4, 11-dien**. *Metab Eng* 2009, **11**:13-19.

19. Hale V, Keasling JD, Renninger N, Diagana TT: **Microbially derived artemisinin: a biotechnology solution to the global problem of access to affordable antimalarial drugs**. *Am J Trop Med Hyg* 2007, **77(Suppl. 6)**:198-202.

20. Martin VJJ, Pitera J, Withers ST, Newman JD, Keasling JD: **Engineering a mevalonate pathway in Eschericia coli for production of terpenoids**. *Nat Chem Biol* 2003, **21**:796-880.

21. Pfleger BF, Pitera DJ, Smolke CD, Keasling JD: **Combinatorial engineering of intergenic regions in operons tunes expression of multiple genes**. *Nat Biotechnol* 2006, **24**:1027-1032.

22. •• Dueber JE, Wu GC, Malmirchegini GR, Moon TS, Petzold CJ, Ullal AV, Prather KLJ, Keasling JD: **Synthetic protein scaffolds provide control over metabolic flux**. *Nat Biotechnol* 2009, **27**:753-759.
'Scaffolding' was used to resolve a bottleneck in the interaction of two heterologous enzymes, hydroxy-methylglutaryl-CoA synthase (HMGS) and HMGR, the downstream reductase. A stoichiometrically programmed enzyme complex was constructed comprising one, two or three SH2 ligands fused to the C-terminal of HMGS, which then captured a corresponding number of HMGRs carrying N-terminal SH2 binding domains. A matrix of different combinations was tested to determine the optimal stoichiometry, substantially increasing pathway flux.

23. Moon TS, Dueber JE, Shiue E, Prather KLJ: **Use of modular, synthetic scaffolds for improved production of glucaric acid in engineered E. coli**. *Metab Eng* 2010, **12**:298-305.

24. •• Ajikumar PK, Xiao W-H, Tyo KEJ, Simeon F, Leonard E, Mucha O, Phon TH, Pfeifer B, Stephanopoulos G: **Isoprenoid pathway optimization for taxol precursor overproduction in Escherichia coli**. *Science* 2010, **330**:70-74.
Reports progress in the challenging reengineering of *E.coli* for production of taxol by an SB pioneer.

25. Gidijala L, Kiel JAKW, Doouma RD, Seifar RM, van Gulik WM, Bovenberg RAL, Veenhuis M, van der Kleil IJ: **An engineered yeast secreting penicillin**. *PLoS ONE* 2009 doi: 10.1371/journal.pone.0008317.g002.

26. Menzella HG, Carney JR, Santi DV: **Rational design and assembly of synthetic trimodular polyketide synthases**. *Chem Biol* 2007, **14**:143-151.

27. •• Walsh C, Fischbach MA: **Natural Products 2.0: connecting genes to molecules**. *J Am Chem Soc* 2010, **132**:2469-2493.
Scholarly, definitive, and readable, this paper is primarily a review of polyketides and non-ribosomal peptides, combined, in the concluding pages, with an appeal for sequencing 1000 secondary metabolite 'workhorse' strains.

28. Ananda S, Prasadi MVR, Yadav G, Kumar N, Shehera J, Ansari Z, Mohanty D: **SBSPKS: structure based sequence analysis of polyketide synthases**. *Nucleic Acids Res* 2010, **38**:W487-W496.

29. Yadev G, Gokhale RS, Mohanty D: **Towards prediction of metabolic products of polyketide synthases: an in silico analysis**. *PLoS Comput Biol* 2009 doi: 10.1371/journal.pcbi.1000351.

30. Sharma J, Gokhale RS, Mohanty D: **In silico analysis of methyltransferase domains involved in biosynthesis of secondary metabolites**. *BMC Bioinform* 2008, **9**:454-475.

31. Marinelli F: **Antibiotics and Streptomyces: the future of antibiotic discovery**. *Microbiol Today* February 2009:21-24.

32. Bentley SD, Chater KF, Cerdeno-Tarraga AM, Challis GL, Thomson NR, James KD, Harris DE, Quail MA, Kieser H, Harper D *et al.*: **Complete genome sequence of the model actinomycete Streptomyces coelicolor A3(2)**. *Nature* 2002, **417**:141-214.

33. Metzker M: **Sequencing technologies – the next generation**. *Nat Genet* 2010, **11**:31-46.

34. • Watve MG, Tickoo R, Jog MM, Bhole BD: **How many antibiotics are produced by the genus Streptomyces?** *Arch Microbiol* 2001, **176**:386-390.
A wonderfully illustrative exercise in statistical model building and refinement. The technical mathematical apparatus is relatively unobtrusive here, while the modeling logic and thought processes are clearly presented.

35. Sommer OA, Church GM, Dantas G: **A functional metagenomic approach for expanding the synthetic biology toolbox for biomass conversion**. *Mol Syst Biol* 2010, **6**:360-368.

36. Schmidt B, Chen C, Liu W-G, Mitchell W: **PheGee@Home: a grid based tool for comparative genomics**. In *Handbook of Research on Computational Grid Technologies for Life Science, Biomedicine and HEalthcare (2 Volumes).* Edited by Mario Cannataro. IGI Global; 2009:168-186.

37. Sorek R, Cossart P: **Prokaryotic transcriptomics: a new view on regulation, physiology and pathogenicity**. *Nat Rev Genet* 2010, **11**:9-16.

38. Li MHT, Ung PMU, Zajkowski J, Garneau-Tsodikova S, Sherman DH: **Automated genome mining for natural products**. *BMC Genomics* 2009, **10**:185-195.

39. Hu Y, Chou WKW, Hopson R, Cane D: **Genome mining in Streptomyces clavuligerus: expression and biochemical characterization of two new cryptic sesquiterpene syntheases**. *Chem Biol* 2011, **18**:32-37.

40. •• Yin T, Zhang X, Gunter L, Priya R, Sykes R, Davis M, Wullschleger SD, Tuskan GA: **Differential detection of genetic loci underlying stem and root lignin content in *Populus***. *PLoS ONE* 2010, **5**:e14.021.
This elegant study employed pyrolysis beam MS to characterize poplar tree stem and root cell wall for 28 lignan components. Lignan composition is a multigenic trait and genes involved were mostly unknown. A full Quantitative Trait Loci analysis was performed to map 'traits' (molecular MS features) to the genome; this led to successful gene discovery. A strong demonstration of phenotype–genotype mapping.

41. Bowen BP, Northen TR: **Dealing with the unknown: metabolomics and metabolite atlases**. *J Am Soc Mass Spectrom* 2010, **21**:1471-1476.

42. Gross H, Stockwell VO, Henkels MD, Nowak-Thompson B, Loper JE, Gerwick WH: **The genomisotropic approach: a systematic method to isolate products of orphan biosynthetic**. *Gene Clusters Chem Biol* 2007, **14**:53-63.

43. Molinaki TF: **NMR of natural products at the 'nanomole-scale'**. *Nat Prod Rep* 2010, **27**:321-329.

44. Schlotterbeck G, Ceccarelli SM: **LC-SPE-NMR-MS: a total analysis system for bioanalysis**. *Bioanalysis* 2011, **3**:549-559.

45. Lin Y, Schiavo S, Orjala J, Vouros P, Kautz R: **Microscale L.C.-M. NMR platform applied to the identification of active cyanobacterial metabolites**. *Anal Chem* 2008, **80**:8045-8054.

46. Yanes O, Tautenhahn R, Patti GJ, Siuzdak G: **Expanding coverage of the metabolome for global metabolite profiling**. *Anal Chem* 2011, **83**:2152-2161.

47. Cheng-Chung L, Snyder TM, Quake SR: **A microfluidic oligonucleotide synthesizer**. *Nucleic Acids Res* 2010, **38**:2514-2521.

48. Gibson DG, Young L, Chuang R-Y, Venter JC, Hutchison CA III, Smith H: **Enzymatic assembly of DNA molecules up to several hundred**. *Nat Methods* 2009, **6**:343-345.

49. Shao Z, Luo Y, Zhao H: **Rapid characterization and engineering of natural product biosynthetic pathways via DNA assembler**. *Mol BioSyst* 2011 doi: 10.1029/cOmb00338g.

50. •• Kosuri S, Eroshenko N, LeProust EM, Super M, Way J, Li JB, Church G: **Scalable gene synthesis by selective amplification of DNApools from high-fidelity microchips**. *Nat Biotechnol* 2010, **28**:1295-1299.
A very important proof of concept paper demonstrating that quality gene assemblies can be created from highly complex pools of cheap, array synthesized oligonucleotides. Superb combination of computation, molecular biology and engineering.

51. Borovkov AY, Loskutov AV, Robida MD, Day KM, Cano JA, Olson TL, Patel H, Brown K, Hunter PD, Sykes KF: **High-quality gene assembly directly from unpurified mixtures of microarray-synthesized oligonucleotides**. *Nucleic Acids Res* 2010 doi: 10.1093/nar/gkg677.

52. •• Gibson DG, Glass JI, Lartigue C, Noskov VN, Chuang R-Y, Algire MA, Benders GA, Chuang R-Y, Montague MG, Ma L *et al.*:

Creation of a bacterial cell controlled by a chemically synthesized genome. *Science* 2010, **329**:52-56.
Beyond the reengineering of metabolic circuitry required for NPSB, this paper reports viable whole genome engineering, the *ne plus ultra*. A landmark.

53. Matzas K, Stahler PF, Kefer N, Siebelt N, Boisguerin V, Leonard JT,
•• Keller A, Stahler CF, Haberle P, Gharizadeh B *et al.*: **High-fidelity gene synthesis by retrieval of sequence-verified DNA identified using high-throughput pyrosequencing**. *Nat Biotechnol* 2010, **28**:1291-1294.
The technique exploits the fact that in NGS a population of identical PCR emulsion cloned sequencing templates is attached to a specific bead and the bead is mapped to a 2-D surface to allow signal detection as the sequencing chemistry proceeds. Therefore, beads of correct and desired sequence, located at a known address, can be physically recovered by micropipette for direct input to gene assembly. Damaged sequences are discarded. In a proof of principle study megacloning resulted in a 500-fold reduction in errors in the final gene assembly as compared to the starting oligonucleotides.

54. Fuhrmann M, Oertel W, Berthhold P, Hegemann P: **Removal of**
• **mismatched bases from synthetic genes by enzymatic mismatch cleavage**. *Nucleic Acids Res* 2005, **33**:e58.
Describes the core technology of ErrASE, which relies on the frequent but random nature of error in oligo synthesis. After PCR amplification, errors, although frequent in aggregate, will be a small percentage of the bases at any particular site, far outnumbered by correct templates. A mismatch endonuclease such as *E.coli* endonuclease V cuts at incorrectly matched basepairs, then a titrated exonuclease activity is deployed to chew back the cut site to remove the erroneous base. After the double stranded population is melted and reannealed, partial strands, those where errors were detected, in general will find perfectly intact hybridization partner strands. Gaps are then filled in with a polymerase to generate healed assemblies.

55. Dixon N, Duncan JN, Geerlings T, Dunstan MS, McCarthy JEG: **Reengineering orthogonally selective riboswitches**. *PNAS* 2010, **107**:2830-2835.

56. Wang HH, Isaacs FJ, Carr PA, Sun ZZ, Xu G, Forest CR,
• Church GM: **Programming cells by multiplex genome engineering and accelerated evolution**. *Nature* 2009, **460**:894-898.
By targeting twenty genes in the *E.coli* lycopene pathway for modification at the ribosomal binding site and simultaneously targeting four genes from competing pathways for disruption, it was possible to increase lycopene yield from 2 mg to 9 mg per gram dry weight in 35 cycles

57. Babiskin AH, Smolke CD: **A synthetic library of RNA control moduled for predictable tuning of gene expression in yeast**. *Mol Syst Biol* 2011, **7**: 101.1038/msb.2011.4.

58. Government of India Ministry of Health and Family Welfare, Department of Ayurveda, Yoga & Naturopathy, Unani, Siddha and Homoeopathy: *The Ayurvedic Pharmacopoeia of India, Vols I–IV*. New Delhi; 2007.

59. Commission of the Ministry of Public Health, PRC: *Pharmacopoeia of the People's Republic of China (English Edition), Vol I, herbal medicine volume*. Beijing; 1997.

60. Shultes, Richard Evans, Raffauf Robert F: *Healing Forest Medicinal and Toxic Plants of Northwest Amazonia*. Dioscorides Press; 2003.