PanDaTox

A tool for accelerated metabolic engineering

Gil Amitai and Rotem Sorek*

Department of Molecular Genetics; Weizmann Institute of Science; Rehovot, Israel

Metabolic engineering is often facilitated by cloning of genes encoding enzymes from various heterologous organisms into *E. coli*. Such engineering efforts are frequently hampered by foreign genes that are toxic to the *E. coli* host. We have developed PanDaTox (www.weizmann.ac.il/pandatox), a web-based resource that provides experimental toxicity information for more than 1.5 million genes from hundreds of different microbial genomes. The toxicity predictions, which were extensively experimentally verified, are based on serial cloning of genes into *E. coli* as part of the Sanger whole genome shotgun sequencing process. PanDaTox can accelerate metabolic engineering projects by allowing researchers to exclude toxic genes from the engineering plan and verify the clonability of selected genes before the actual metabolic engineering experiments are conducted.

Metabolic Engineering and the Problem of Toxic Genes

Metabolic engineering is a rapidly growing field where enzymatic pathways for the biosynthesis of desired molecules are genetically engineered into model microorganisms in order to harness bacterial productivity into industrial use.[1] Metabolic engineering is often practiced in close connection with synthetic biology and systems biology, as significant theoretical modeling is conducted in order to model pathways to be engineered into a given organism.[2-5]

The bacterial species *Escherichia coli* is one of the most widely used microorganisms in metabolic engineering and had been utilized for numerous bio-production applications. For example, massive production of precursors for the antimalarial drug artimisinin was facilitated by inserting genes from several microorganisms into *E. coli*;[6,7] polyketides, which are the precursors for many antibiotics, have been produced within *E. coli* by introducing a combination of genes from three different bacteria into this organism;[8] various nutritional molecules such as acetate, pyruvate, succinate and an array of amino acids are also among the products produced within *E. coli* through metabolic engineering.[9] Recently, metabolic engineering has taken center stage in the global efforts for the design and generation of biofuel-producing organisms, in attempts to generate biological alternatives to fossil fuels.[1,10]

Thousands of microbial genomes have been sequenced to date[5,11] and these genomes cumulatively harbor millions of enzymes that can be used as "building blocks" in the toolbox of the metabolic engineer. Following detailed planning of the pathway of new enzymes to engineer into *E. coli*, the metabolic engineer will usually search for these enzymes in the set of available genomes and will select genes encoding the desired enzymes from organisms in which these genes exist. The selected genes will then be cloned into *E. coli* with or without optimization for expression through codon and promoter alterations.[3,12]

One of the major hurdles in such metabolic engineering efforts is genes that are toxic to the receiving organism.[4,6] Many enzymes are toxic to *E. coli* when cloned

into this organism heterologously and their toxicity may stem from generation of toxic intermediates or other incompatibility issues.[13] Attempts to engineer a toxic gene into *E. coli* will usually fail to produce viable clones and will significantly slow down the engineering project. Furthermore, since the experimental procedures in metabolic engineering projects are significantly longer and more expensive as compared with the theoretical design, months of effort and considerable funds expenditure could be wasted due to toxic genes.

PanDaTox: The Pan Genomic Database of Genes Toxic to Bacteria

We have recently introduced PanDaTox, the pan genomic database of genes and genomic elements that are toxic to bacteria (www.weizmann.ac.il/pandatox).[13] PanDaTox contains experimental toxicity information for over 1.5 million genes from hundreds of microorganisms and reports the results of cloning attempts for each of these genes on single-copy and multiple-copy vectors within *E. coli* (**Fig. 1**). This resource is expected to accelerate metabolic engineering efforts by allowing researchers to perform a-priori exclusion of toxic genes from the engineering plan, and on the other hand, a-priori verification of clonability of selected genes before the actual experiments.

Although our identification of toxic genes was based on massive experimental cloning of over 9 million clones into *E. coli*, this was not an intentionally-designed experiment. In fact, the detection of toxic genes was a by-product of the Sanger whole-genome sequencing procedure through which hundreds of microbial genomes had been sequenced. Within this procedure, multiple copies of the sequenced genome are randomly fragmented into overlapping pieces of DNA that are serially inserted into *E. coli* on plasmids. These fragments are then sequenced and assembled based on sequence overlaps. While usually most fragments can be cloned into the *E. coli* host, a small fraction of the genome fails to be cloned and this results in sequence gaps that interfere with proper genome assembly (**Fig. 2**).

In an earlier study as well as in our recent study (references 13 and 14), we have demonstrated that gaps in microbial genome sequences are frequently caused by genes that are toxic to the *E. coli* host. These genes inhibit *E. coli* growth when cloned into it and are hence missing from the set of sequences available for genome assembly. Following gap closure (which is performed by cloning-independent methods), these toxic genes can readily be identified. By experimenting with cumulatively more than 100 gap-residing genes from numerous different organisms, we have verified that our gap-based prediction of gene toxicity exceeds 80% of accuracy.[13,14]

The accuracy in our predictions of gene clonability into *E. coli* stems from several factors. First, the high sequencing coverage, needed for proper genome assembly, dictates that on average more than 25 independent clones would contain each gene. This ensures that the absence of a specific gene from the set of sequencing clones is not likely to occur by chance only and provides high statistical power for identification of toxic genes. Furthermore, the diversity of cloning libraries used for genome sequencing, where some clones are propagated on a high-copy number plasmid while others are found on single-copy plasmid allows us to differentiate between stronger and weaker toxicity, i.e., differentiate between genes that will be toxic only when highly expressed and those which are toxic even in low doses.

PanDaTox contains data on over 40,000 genes that are predicted to be toxic to *E. coli* based on their clonability. It allows text-based searches for enzymes of interest according to multiple keywords, as well as sequence-based searches (blast) with a user-provided gene; homology searches enable checking whether homologs of a selected toxic gene are also toxic when cloned from a different genome; and clone coverage analysis allows the user to evaluate the toxicity inference. PanDaTox presents toxicity information in both graphical and numerical manners and provides links to the experimental results collected.

With the advent of cheap genome sequencing, sequences of whole microbial genomes accumulate at a rapid pace.

These, in turn, enrich the toolbox of metabolic engineers with numerous potential enzymes and pathways, but also increase the challenge in selecting the proper genes from the proper organisms. PanDaTox, which contains toxicity information for over 1.5 million microbial genes, can accelerate metabolic engineering by helping scientists to avoid toxic genes and to select genes that were experimentally proven to be clonable within *E. coli*. With the rapidly growing interest in metabolic engineering both in the academy and in the industry, we expect PanDaTox to form a useful tool for the community.

References

1. Keasling JD. Manufacturing molecules through metabolic engineering. Science 2010; 330:1355-8; http://dx.doi.org/10.1126/science.1193990.

2. Keasling JD. Synthetic biology for synthetic chemistry. ACS Chem Biol 2008; 3:64-76; http://dx.doi.org/10.1021/cb7002434.

3. Nielsen J, Keasling JD. Synergies between synthetic biology and metabolic engineering. Nat Biotechnol 2011; 29:693-5; http://dx.doi.org/10.1038/nbt.1937.

4. Keasling JD. Synthetic biology and the development of tools for metabolic engineering. Metab Eng 2012; 14:189–95; http://dx.doi.org/10.1016/j.ymben.2012.01.004.

5. Sorek R, Serrano L. Bacterial genomes: from regulatory complexity to engineering. Curr Opin Microbiol 2011; 14:577-8; http://dx.doi.org/10.1016/j.mib.2011.09.006.

6. Martin VJ, Pitera DJ, Withers ST, Newman JD, Keasling JD. Engineering a mevalonate pathway in *Escherichia coli* for production of terpenoids. Nat Biotechnol 2003; 21:796-802; http://dx.doi.org/10.1038/nbt833.

7. Newman JD, Marshall J, Chang M, Nowroozi F, Paradise E, Pitera D, et al. High-level production of amorpha-4,11-diene in a two-phase partitioning bioreactor of metabolically engineered *Escherichia coli*. Biotechnol Bioeng 2006; 95:684-91; http://dx.doi.org/10.1002/bit.21017.

8. Pfeifer BA, Admiraal SJ, Gramajo H, Cane DE, Khosla C. Biosynthesis of complex polyketides in a metabolically engineered strain of *E. coli*. Science 2001; 291:1790-2; http://dx.doi.org/10.1126/science.1058092.

9. Wendisch VF, Bott M, Eikmanns BJ. Metabolic engineering of *Escherichia coli* and Corynebacterium glutamicum for biotechnological production of organic acids and amino acids. Curr Opin Microbiol 2006; 9:268-74; http://dx.doi.org/10.1016/j.mib.2006.03.001.

10. Lee SK, Chou H, Ham TS, Lee TS and Keasling JD. Metabolic engineering of microorganisms for biofuels production: from bugs to synthetic biology to fuels. Curr Opin Biotechnol 2008; 19:556-63; http://dx.doi.org/10.1016/j.copbio.2008.10.014.
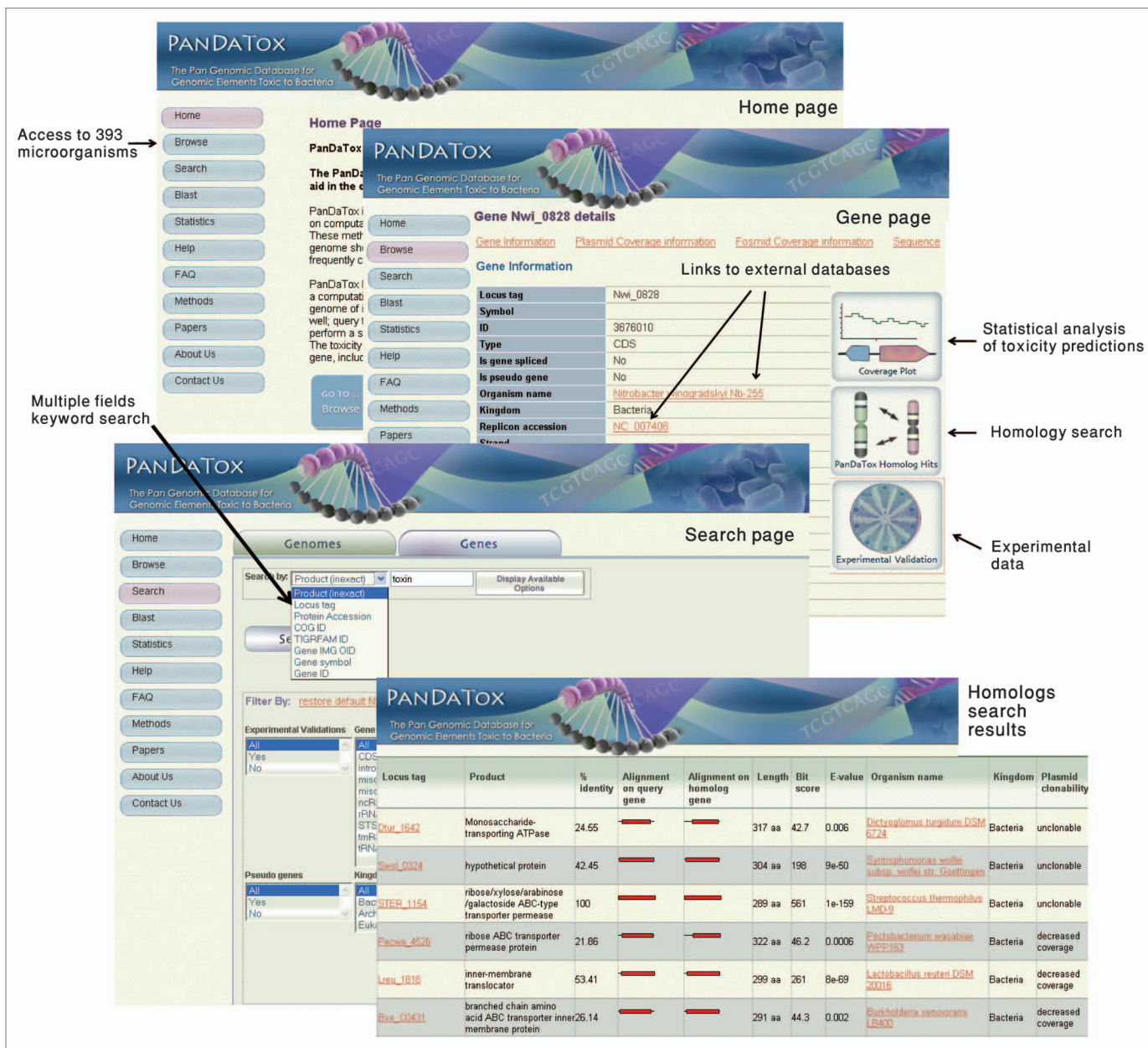
Figure 1. The PanDaTox online web tool. PanDaTox holds toxicity information for over 1.5 million genes. It presents detailed information for genes of choice; allows searching by multiple keywords; and provides links to multiple external sources. The home page provides access to the database; the gene page holds multiple clickable details on gene annotations, sequences, toxicity analyses and presents results of experiments done with these genes (when applicable); the search page enables searching for genes of interest by keywords and various filters; the homologs page presents toxicity of homologs from other genomes. Users can also search for their genes of interest by sequence-based search (using the Blast page).

11. Pagani I, Liolios K, Jansson J, Chen IM, Smirnova T, Nosrat B, et al. The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. Nucleic Acids Res 2012; 40:571-9; http://dx.doi.org/10.1093/nar/gkr1100.

12. Boyle NR, Gill RT. Tools for genome-wide strain design and construction. Curr Opin Biotechnol 2012; In Press; http://dx.doi.org/10.1016/j.copbio.2012.01.012.

13. Kimelman A, Levy A, Sberro H, Kidron S, Leavitt A, Amitai G, et al. A vast collection of microbial genes that are toxic to bacteria. Genome Res 2012; 22:802–9; http://dx.doi.org/10.1101/gr.133850.111.

14. Sorek R, Zhu Y, Creevey CJ, Francino MP, Bork P, Rubin EM. Genome-wide experimental determination of barriers to horizontal gene transfer. Science 2007; 318:1449-52; http://dx.doi.org/10.1126/science.1147112.
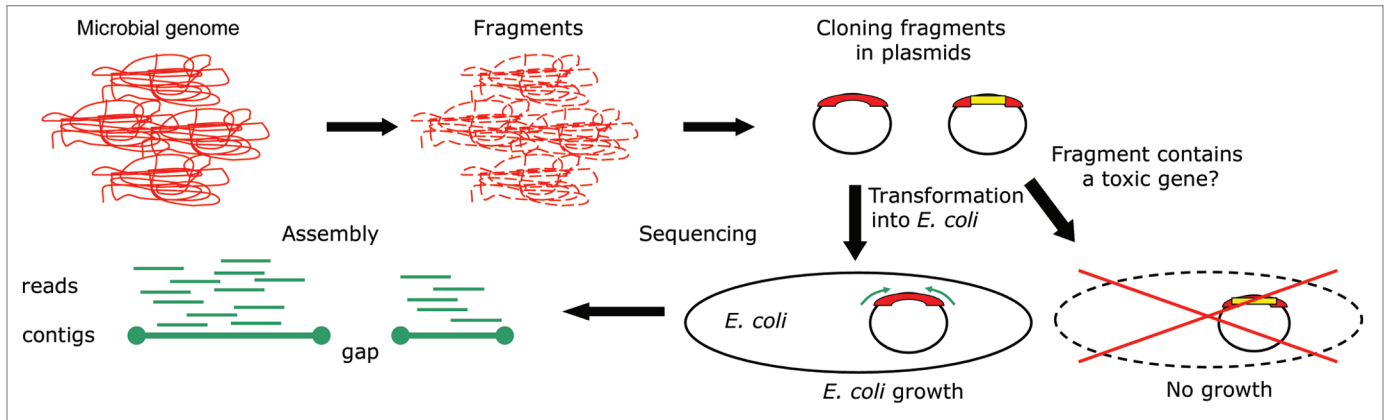
Figure 2. Genes that are toxic to *E. coli* are exposed as a byproduct of the microbial genome sequencing process. The sequenced genome is physically sheared into overlapping fragments of DNA, which are transformed into *E. coli* bacteria. Fragments are sequenced and assembled into contigs. Toxic genes result in *E. coli* growth inhibition and are hence not properly sequenced, creating "gaps" in genome assemblies. After gap closure the gap sequences are retrieved. Post-analysis of gap-residing genes exposes the toxic genes.