Journal of Biotechnology xxx (2013) xxx–xxx

Contents lists available at ScienceDirect

Journal of Biotechnology

journal homepage: www.elsevier.com/locate/jbiotec

Biomining active cellulases from a mining bioremediation system☆

Keith Mewis[a], Zachary Armstrong[a], Young C. Song[b], Susan A. Baldwin[c], Stephen G. Withers[a,d,e], Steven J. Hallam[a,b,f,*]

[a] Genome Science and Technology Program, Canada
[b] UBC Graduate Program in Bioinformatics, Canada
[c] Department of Chemical and Biological Engineering, Canada
[d] Department of Chemistry, Canada
[e] Centre for High Throughput Biology, Canada
[f] Department of Microbiology & Immunology, Canada

ARTICLE INFO

Article history:
Received 1 May 2013
Received in revised form 8 July 2013
Accepted 11 July 2013
Available online xxx

Keywords:
Functional screening
Metagenomics
Cellulase
Mining bioreactor

ABSTRACT

Functional metagenomics has emerged as a powerful method for gene model validation and enzyme discovery from natural and human engineered ecosystems. Here we report development of a high-throughput functional metagenomic screen incorporating bioinformatic and biochemical analyses features. A fosmid library containing 6144 clones sourced from a mining bioremediation system was screened for cellulase activity using 2,4-dinitrophenyl β-cellobioside, a previously proven cellulose model substrate. Fifteen active clones were recovered and fully sequenced revealing 9 unique clones with the ability to hydrolyse 1,4-β-D-glucosidic linkages. Transposon mutagenesis identified genes belonging to glycoside hydrolase (GH) 1, 3, or 5 as necessary for mediating this activity. Reference trees for GH 1, 3, and 5 families were generated from sequences in the CAZy database for automated phylogenetic analysis of fosmid end and active clone sequences revealing known and novel cellulase encoding genes. Active cellulase genes recovered in functional screens were subcloned into inducible high copy plasmids, expressed and purified to determine enzymatic properties including thermostability, pH optima, and substrate specificity. The workflow described here provides a general paradigm for recovery and characterization of microbially derived genes and gene products based on genetic logic and contemporary screening technologies developed for model organismal systems.

1. Introduction

Cellulose, one of the most abundant sources of organic carbon on the planet, has wide-ranging industrial applications, with increasing emphasis on biofuel production. As a result, many studies have extensively focused on the identification of carbohydrate active enzymes, or CAZymes, using both culture-dependent (Rastogi et al., 2009) and culture-independent (Xia et al., 2013) methods. The CAZy database (http://www.cazy.org) currently defines 131 families of glycoside hydrolases (GHs) based on sequence and structure providing a useful resource for functional annotation of predicted GH genes (Cantarel et al., 2009). Seventeen of these families are reported to have cellulase activity, classified by their ability to hydrolyse 1,4-β-D-glucosidic linkages found in cellulose, lichenan, and cereal β-D-glucans. The current production of cellulosic ethanol from non-feedstock crops typically utilizes enzymatic hydrolysis steps to break cellulose into its constituent sugars prior to fermentation (Brethauer and Wyman, 2010). The current high cost of versatile industrial enzymes is a limiting factor in this production (Lee et al., 2010), necessitating the discovery or development of new enzymes that may show more desirable attributes conducive to current cellulosic ethanol pipelines, such as improved acid and temperature stability. Many organisms have been enriched and cultured with this intention, but the biggest reservoir of microbial diversity remains uncultured and untapped within natural and human engineered ecosystems.

To address this cultivation gap, functional metagenomic screens have been developed to recover active genes sourced directly from environmental samples (reviewed in Taupp et al. (2011)). While the discovery of many different enzyme classes has been reported, cellulases have been among the most sought after genes from a biotechnological perspective. Functional metagenomic screens to

* Corresponding author at: University of British Columbia, Department of Microbiology & Immunology, Life Sciences Institute, 2552-2350 Health Sciences Mall, Vancouver, British Columbia V6T1Z3, Canada. Tel.: +1 604 827 3420; fax: +1 604 822 6041.
E-mail address: shallam@interchange.ubc.ca (S.J. Hallam).

0168-1656/$ – see front matter © 2013 The Authors. Published by Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jbiotec.2013.07.015

identify novel cellulases have been conducted on environmental samples from soils (Nacke et al., 2012; Voget et al., 2006), gut microbiomes (Pope et al., 2010; Warnecke et al., 2007) and a biogas plant (Ilmberger et al., 2012). Previously we reported a method using a 384-well plate format to allow for increased throughput and scalability of liquid-assay screens (Mewis et al., 2011). This format allows for the incorporation of automated liquid handling systems and a quantitative readout for accurate comparison between samples.

Here we build upon this automated screening paradigm to recover and characterize active cellulase-encoding fosmid clones sourced from a biochemical reactor (BCR) system designed for remediation of metal contaminated water. The bioreactor used a cellulosic substrate or feedstock composed mostly of pulp mill biosolids to reduce or remove arsenic, cadmium, and zinc from smelter waste seepage through both biotic and abiotic processes (Kawaja et al., 2005; Mattes et al., 2011). We present the development and application of an automated tree-building pipeline for phylogenetic assignment of discovered genes within GH 1, 3 and 5 families and demonstrate an efficient purification and characterization process to constrain the biochemical properties of active clones for biomass transformation within the bioreactor system.

2. Materials and methods

2.1. Fosmid library construction

A fosmid library was constructed using high molecular weight (HMW) DNA extracted from a homogenized core sample derived from a BCR (site details in Supp. Info). This bioreactor operates year round at temperatures ranging from 0 °C to 18 °C from August 2008 to July 2009 when the reactor was sampled. Environmental DNA was cloned into the pCC1 copy control system using *Escherichia coli* EPI300 as expression host (Epicentre, Madison, WI) as previously described (Taupp et al., 2009). Library construction yielded 6144 fosmid harbouring clones with average insert size of 42 kilobase (kb) pairs. Bi-directional Sanger end-sequencing was performed using the ABI BigDye kit (Applied Biosystems, Carlsbad, Ca) on all clones at Canada's Michael Smith Genome Sciences Centre, Vancouver, BC, Canada with the pCC1 forward (5′-GGATGTGCTGCAAGGCGATTAAGTTGG) and reverse (5′-CTCGTATGTTGTGTGGAATTGTGAGC) sequencing primer.

2.2. High throughput functional screening

Cellulase activities were assayed in 384-well culture plates using 2,4-dinitrophenyl β-cellobioside (gift from Dr. Hongming Chen, Withers Lab, UBC) as reported previously (Mewis et al., 2011). The DNP group is a desirable chromophore for this application due to its lower pK_a value than the commonly used p-nitrophenol (pNP) chromophore, allowing the assay to be conducted at full sensitivity at the environmental pH of 6.9. The assay temperature was chosen as 37 °C as this was the optimal temperature for *E. coli* growth, and hence the temperature at which the enzymes were expressed. Wells were considered cellulase active if they exhibited 400 nm absorbance more than six standard deviations above the sample mean of all plates. A 96-well glycerol stock plate containing six copies of each active clone was created, and a copy of this plate was screened in the same manner to ensure there were no location or volume effects contributing to the hydrolysis of substrate. Screening of this plate enabled a quantitative comparison between the activities of individual clones without solution or plate-location heterogeneity.

2.3. Full fosmid sequencing

Once active clones were identified, fosmid DNA was extracted using the FosmidMax DNA preparation kit (Epicentre) according to the manufacturer's instructions, and further treated with PlasmidSafe DNase (Epicentre) to remove contamination *E. coli* chromosomal DNA. DNA concentrations were measured using Quant-iT PicoGreen (Invitrogen, Carlsbad, CA). For full fosmid sequencing, 500 ng of each fosmid was sent to Canada's Michael Smith Genome Sciences Centre (Vancouver, BC). In order to maximize sequencing throughput, 92 individual fosmid samples (not all reported here) were barcoded and sequenced on a single lane of an Illumina GAIIx sequencer (Illumina, San Diego, CA). Contigs were assembled for each well using the barcoded sequences with ABySS v1.2 (Simpson et al., 2009).

2.4. Transposon mutagenesis

For each active clone, a Tn5 transposon mutagenesis library was created (EZ-Tn5 *kan* insertion kit, Epicentre) to identify which gene on the insert was responsible for detected activity. The region surrounding the insertion was sequenced using an automated DNA sequencer (Applied Biosystems 3730 system, Carlsbad, CA) and primers complementary to the inserted transposon. Sequence data was assembled using phred/phrap (http://www.phrap.org/phredphrapconsed.html) (Ewing et al., 1998), and Consed was used to examine and export resulting contigs into fasta file format (http://www.phrap.org/consed/consed.html) (Gordon et al., 1998). Assembled contigs from these Sanger sequences ranged in size from 382 bp to 5312 bp in length.

A custom perl script was designed that uses Circos (Krzywinski et al., 2009) to visualize contig relationships between transposon contigs and full sequences. This was necessary to remove "hitchhiking" sequences originating from other wells during the library construction process. Transposon data for one well (F1) localized to contigs from two separate wells likely due to contamination during transposon mutagenesis and thus transposon insertion information from this fosmid was limited.

2.5. Gene finding and open reading frame prediction

Open reading frames (ORFs) were determined for end sequences, full fosmid sequences and transposon contigs using the metagenome option of Prodigal (http://prodigal.ornl.gov/) (Hyatt et al., 2010). End sequences yielded 17,648 predicted ORFs, full fosmid sequencing yielded 623 predicted ORFs and transposon mutagenesis contigs yielded 169 predicted ORFs. These ORFs were compared to the CAZy protein database using BLASTP (Altschul et al., 1997) with an expectation value cutoff of $1 \times e^{-20}$ using criteria reported by Martinez et al. (2010).

2.6. Generating reference datasets for GH families 1, 3, and 5

All protein sequences from GH 1 (1630 sequences), GH 3 (2297 sequences), and GH 5 (875 sequences) families were downloaded from the CAZy database in August 2011. Sequences were filtered by length (see supplementary methods) and clustered with UCLUST (Edgar, 2010), representative sequences were aligned with MUSCLE (Edgar, 2004), and then inserted into RAxML (Stamatakis, 2006). Hidden Markov models were generated using *hmmbuild* (Eddy, 1998). Complete reference datasets include alignment, hidden markov model, and the phylogenetic tree. Following the generation of reference data sets, each family was appended to the MLTreeMap package as a set of functional marker genes. The phylogeny of identified GH 1, GH 3 and GH 5 genes was

Please cite this article in press as: Mewis, K., et al., Biomining active cellulases from a mining bioremediation system. J. Biotechnol. (2013), http://dx.doi.org/10.1016/j.jbiotec.2013.07.015