ELSEVIER

FEMS Microbiology Letters 211 (2002) 175–182

FEMS MICROBIOLOGY Letters

www.fems-microbiology.org

Use of subtractive hybridization for comprehensive surveys of prokaryotic genome differences

Peter G. Agron [a], Madison Macht [a], Lyndsay Radnedge [a], Evan W. Skowronski [a,1], Webb Miller [b], Gary L. Andersen [a,*]

[a] *Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, L-441, 7000 East Avenue, Livermore, CA 94551, USA*
[b] *Department of Computer Science and Engineering, Pennsylvania State University, University Park, PA 16802, USA*

Received 7 March 2002; received in revised form 2 April 2002; accepted 2 April 2002

First published online 25 April 2002

Abstract

Comparative bacterial genomics shows that even different isolates of the same bacterial species can vary significantly in gene content. An effective means to survey differences across whole genomes would be highly advantageous for understanding this variation. Here we show that suppression subtractive hybridization (SSH) provides high, representative coverage of regions that differ between similar genomes. Using *Helicobacter pylori* strains 26695 and J99 as a model, SSH identified approximately 95% of the unique open reading frames in each strain, showing that the approach is effective. Furthermore, combining data from parallel SSH experiments using different restriction enzymes significantly increased coverage compared to using a single enzyme. These results suggest a powerful approach for assessing genome differences among closely related strains when one member of the group has been completely sequenced. © 2002 Published by Elsevier Science B.V. on behalf of the Federation of European Microbiological Societies.

Keywords: Bacterial genome; Comparative genomics; Subtractive hybridization; *Helicobacter pylori*

1. Introduction

Subtractive hybridization has become an increasingly important tool for comparative prokaryotic genomics, because it allows significant enrichment of sequences that differ between two DNA samples. Phenotypic variation in prokaryotes is often associated with differences in gene content, so the isolation of DNA fragments from these genes or associated regions is illuminating. Examples include the identification of virulence genes in *Burkholderia mallei* [1], pathogen-specific genes in *Neisseria* [2], serovar-specific genes in *Salmonella enterica* [3], and novel restriction-modification systems in *Helicobacter pylori* [4]. Even if the function of the different regions is unclear, the nucleotide sequences can serve as diagnostic markers for DNA-based detection (e.g. [5–8]).

When a complete genome sequence is available, microarray technology using DNA–DNA hybridization can be useful for defining regions that are present in the sequenced strain, but absent in unsequenced strains [9–12]. To define unique regions present in an unsequenced strain, one option is to completely sequence the close relative, then perform computer sequence comparisons. This approach has proven fruitful in several cases, but the projects required substantial effort and there was little prior knowledge about the nature or amount of differences likely to be found. For example, a comparison between the complete genomes of *Escherichia coli* K-12, a non-pathogenic strain, and *E. coli* EDL-933, an enterohemorrhagic strain, provides a genetic basis for the different lifestyles of the two strains [13,14]. Two pathogenic strains of *H. pylori* have also been sequenced, and show significant differences in gene content [15,16]. In both examples a predominant source of genomic variation is the insertion or deletion

* Corresponding author. Tel.: +1 (925) 423-2525; Fax: +1 (925) 422-2282.
E-mail address: andersen2@llnl.gov (G.L. Andersen).

[1] Present address: MJ GeneWorks, 7000 Shoreline Court, South San Francisco, CA 94080, USA.

Abbreviations: SSH, suppression subtractive hybridization; ORF, open reading frame; PIP, percent identity plot

0378-1097/02/$22.00 © 2002 Published by Elsevier Science B.V. on behalf of the Federation of European Microbiological Societies.
PII: S0378-1097(02)00650-X

of large (10–50-kb) DNA regions, which subtractive hybridization can easily identify. These studies underscore the need for sequence information from more than one isolate within a closely related group. Given the effort required for whole genome sequencing projects, it would be highly advantageous to have an effective means to examine candidate strains for differences, thus allowing informed choices for further studies.

Subtractive hybridization may provide a means to perform comprehensive surveys of genome differences among closely related strains. Sequence information from comprehensive surveys will provide valuable insights into biological differences, and can be used to assess candidate strains for their suitability for additional comparative genomic studies. Subtractive hybridization facilitates the identification of DNA segments present in one genome, termed the tester, but absent in another, termed the driver. If a complete sequence is available for the driver, false positives can be easily identified by computer sequence comparisons and discarded, making the approach extremely powerful. Several different subtractive hybridization methods have been developed (e.g. [17–20]), but all are based on the melting and reannealing of a mixture of tester and driver DNA fragments followed by the enrichment of tester-specific sequences. In most of these methods, including suppression subtractive hybridization (SSH), tester and driver fragments are generated by digestion with a restriction endonuclease [18]. SSH has the advantage that it is PCR-based and does not require physical separations.

We have examined the efficacy of SSH to survey prokaryotic genomes for differences in gene content. While subtractive hybridization has previously been shown to be useful for defining a handful of differences between bacterial genomes, the ability of this technique to identify nearly all differences in gene content has not been tested. SSH using two sequenced strains of *H. pylori* provided an ideal model to demonstrate high, representative genome coverage. The results presented here are readily applicable to situations in which there is little or no sequence information for the tester strain. We also show that using multiple restriction enzymes in parallel SSH experiments significantly reduces the effort required to reach high coverage and specificity.

2. Materials and methods

2.1. DNA preparation

H. pylori strains 26695 and J99 were obtained, respectively, from Tim McDaniel, Stanford University, Stanford, CA, USA and Richard Alm, AstraZeneca Corporation, Boston, MA, USA. Bacteria were grown on horse blood agar under microaerobic conditions at 37°C. Blood agar contained 4.4% Columbia agar base (Difco), 5% defibrinated horse blood (HemoStat, Dixon, CA, USA), 10 μg ml^{-1} vancomycin, 5 μg ml^{-1} cefsulodin, 2.5 U ml^{-1} polymyxin B, 50 μg ml^{-1} cycloheximide, 5 μg ml^{-1} trimethoprim, 8 μg ml^{-1} amphotericin B, and 0.2% β-cyclodextrin (all antimicrobials from Sigma). Microaerobic conditions were established using a BBL bell jar with the BBL CampyPak (Becton Dickinson) generation system. Cells from two to four 9-cm plates were collected and washed three times in TE+NaCl (10 mM Tris–HCl pH 8.0, 1 mM EDTA pH 8.0, 150 mM NaCl). The cell pellet was freeze/thawed once using a dry ice/ethanol bath, then DNA was extracted with 1 ml DNA STAT-60 (Tel-Test, Friendswood, TX, USA) according to the manufacturer's instructions. DNA was dissolved in TE to a concentration of approximately 0.3 μg ml^{-1}. Restriction enzymes (New England Biolabs, Beverly, MA, USA) were used according to the manufacturer's instructions. After digestion and heat inactivation, DNA was concentrated by ethanol precipitation in the presence of 20–30 μg ml^{-1} linear acrylamide as a carrier (Ambion, Austin, TX, USA) and dissolved in water at 300 ng μl^{-1}.

2.2. Suppression subtractive hybridization

Genome comparisons by SSH were performed as previously described except for the differences noted below [18]. Briefly, the tester DNA is digested with a restriction endonuclease and the fragments are marked by ligation to specialized oligonucleotide adapters. When the marked DNA is denatured and hybridized to excess unmarked driver DNA that has been digested with the same enzyme, most tester sequences will form heterohybrids with the driver. Tester-specific sequences, however, will self-hybridize to form amplifiable fragments that are then enriched by PCR, cloned, and sequenced. Modified adapters were constructed to allow for subtractions with *Sau*3AI-digested DNA: adapter 1 was formed by annealing the adapter 1 long oligonucleotide [18] with the oligonucleotide 5′-GATCACCTGCCCGG to form an adapter with appropriate cohesive ends. Adapter 2 was formed by annealing the adapter 2 long oligonucleotide with the oligonucleotide 5′-GATCCAATCGGCCG. Similarly, for *Apo*I-digested DNA, adapter 1 was formed using the oligonucleotide 5′-AATTACCTGCCCGG, while adapter 2 was formed using the oligonucleotide 5′-AATTCAATCGGCCG. T4 DNA ligase (New England Biolabs) was inactivated by incubation at 72°C for 20 min.

2.3. DNA sequencing

Unpurified PCR products were cloned using the pGEM T-Easy TA cloning kit (Promega, Madison, WI, USA). Recombinant clones were picked using a BioRobotics (Woburn, MA, USA) BioPick automated colony picker, and plasmid templates were prepared by boiling lysis and magnetic bead capture using a high throughput procedure

[21]. Sequencing of plasmid templates was performed using the Applied Biosystems (Foster City, CA, USA) BigDye Terminator system and either ABI 377 or 3700 automated sequencers. The sequencing primers used were 5′-TGTAAAACGACGGCCAGT (forward) and 5′-CAGGAAACAGCTATGACC (reverse).

2.4. Sequence analysis

Sequences were examined in batch by comparisons to both *H. pylori* J99 and 26695 sequences using the cross_match program, part of the Consed software package [22]. The sequences for each genome were downloaded from GenBank for this purpose, using accession numbers AE001439 and AE000511 for strains J99 and 26695, respectively. Cross_match allows large amounts of data to be processed quickly in parallel on a local UNIX platform. The output was downloaded to Microsoft Excel for further manipulations. The cross_match program [22] assigns a similarity score to each subtraction product when compared to a designated sequence file. To calculate a score, the optimal alignment is found, and each position is assigned a value of +1 for a match, −2 for a mismatch, −2 for the initiation of a gap, and −1 for every additional gap position. For example, a 400-bp fragment will have a score of 400 if there is an exact match to the driver genome. If no statistically significant alignment to the driver genome is found, the cross_match program defines the score as 0. To allow direct comparisons among fragments by normalizing scores, the cross_match score was divided by the length of the sequence used for comparison. In the case of a perfect match, the normalized cross_match score is 1, indicating the region is identical in both genomes, i.e. a false positive. Less than perfect matches yield lower cross_match scores, which when normalized yield values between 0 and 1. Negative cross_match scores are not obtained, as sequences with no or poor matches do not yield an alignment. To determine the degree of similarity necessary for subtraction by SSH, sequence data were compared from subtracted versus unsubtracted clones (see Section 3). To generate unsubtracted clones, strain 26695 genomic DNA was digested with *Alu*I and ligated to both adapters 1 and 2 in the same reaction, whereupon it was amplified, cloned, and sequenced as in the SSH experiments.

3. Results

3.1. Identification of strain-specific DNA regions

H. pylori provides an excellent model system to examine the ability of subtractive hybridization to provide high coverage information about strain-specific gene content. Two strains, J99 and 26695, have been completely sequenced and show approximately 7% unique open reading frames (ORF) for each strain. The complete sequence of each strain allows straightforward identification of false positives and also allows the degree of genome coverage for the tester strain to be easily assessed.

Our first step toward using SSH for strain comparisons was to determine the degree of difference between tester and driver fragments that is required for enrichment by this method. This was examined by comparing sequences obtained from SSH with sequences obtained from unsubtracted fragments. Cross_match [22] was used for batch comparisons to the driver genome (strain J99), and the scores were normalized as described in Section 2. The numbers of fragments falling in given score ranges were counted and the data for subtracted and unsubtracted clones were compared. It was observed that enriched fragments, i.e. those whose numbers increased with SSH, had normalized scores of 0 to approximately 0.2, while the number of other sequences decreased proportionately (data not shown). A score of 0 means there is no similarity, while a score of 1 indicates the two sequences are identical. To our knowledge, a systematic examination of the degree of similarity necessary for subtractive hybridization has not been previously performed. Based on these data, we define tester-specific sequences (difference products) as those falling in this score range (0–0.2) when compared to the driver genome.

Experimental subtractions were then performed to determine if SSH is capable of comprehensively identifying dissimilar regions across a whole genome. Table 1 shows that for each strain, four SSH experiments using different restriction enzymes (*Alu*I, *Apo*I, *Dra*I and *Sau*3AI) were performed. This resulted in the sequencing of 3264 *H. pylori* J99 clones (using 26695 as the driver) and 3744 *H. pylori* 26695 clones (using J99 as the driver, Table 1a). For the J99 and 26695 sequences the average pass rates were 90% and 86%, respectively (Table 1b). Of these high

Table 1
Summary of DNA sequencing results

Restriction enzyme used for SSH	Total number of clones sequenced		Sequencing pass rate (%)		Tester-specific sequences (%)	
	J99	26695	J99	26695	J99	26695
*Alu*I	864	960	92	94	22	30
*Apo*I	864	1248	92	84	11	17
*Dra*I	768	768	87	85	25	19
*Sau*3AI	768	768	88	81	17	25
Total	3264	3744	90	86	19	23

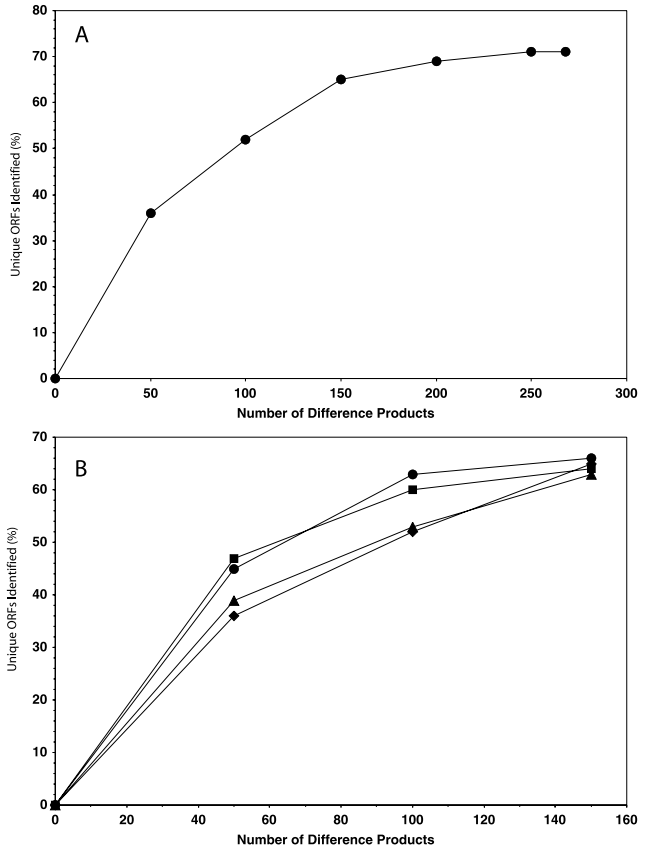

Fig. 1. A: Percentage of unique ORFs identified by SSH with increasing numbers of difference products isolated using a single restriction enzyme *Alu*I. B: Percentage of unique ORFs identified by SSH with increasing numbers of difference products isolated using four restriction enzymes: *Alu*I (diamond), *Apo*I (square), *Dra*I (circle) and *Sau*3AI (triangle).

quality sequences, the proportion of difference products for a given experiment varied from 11% to 30% (Table 1c). These differences may reflect experiment to experiment variation, but may also indicate the ability of a given restriction enzyme to generate optimal size fragments within each of the difference regions. Control experiments in which unsubtracted clones were sequenced revealed 2–3% difference products, showing that substantial enrichment was achieved. Parallel, independent subtractions using *Alu*I-digested DNA and strain 26695 as the tester showed 30% positives in each case indicating that the procedure is reproducible (data not shown).

Difference products were then mapped to the tester genome also using cross_match, and compared to the coordinates of strain-specific ORFs identified by Alm et al. [15]. An ORF was designated as 'identified' if a subtraction product mapped within 1 kb of the ORF. Using this criterion, genome-wide coverage was clearly demonstrated. Seven of the 117 unique ORFs reported by Alm et al. [15] were excluded from this analysis because the ORFs are interrupted in one strain, but not the other, which precludes identification by subtractive hybridization. For experiments using *H. pylori* 26695 as the tester, all but five of the remaining ORFs were identified: 6HP 670, 6HP 967, 6HP 986, 6HP 1188, and 6HP 1404. Given that 110 ORFs are *H. pylori* 26695-specific, this means 95% were identified. The 6HP 967 ORF is 285 bp and 6HP 1404 is 288 bp, suggesting the possibility that the small size of these ORFs prevented their identification. The five missing ORFs are distributed throughout the genome, implying that there are no areas of the genome refractory to subtraction. Analogous results were obtained when strain *H. pylori* J99 was used as the tester, showing the consistency of the approach. Of the 89 *H. pylori* J99-specific ORFs, only five were not identified (94% coverage): J99 331, J99 332, J99 726, J99 940, and J99 1306. ORFs J99 331 and J99 332 are separated by 34 bp, and are 294 and 258 bp in length, respectively. Taken together, these results indicate that high coverage isolation of unique fragments was obtained by subtractive hybridization, and show that this is a viable approach for high confidence strain comparisons.

3.2. Different restriction enzymes provide similar coverage

SSH is a PCR-based technique, so large fragments are not efficiently amplified so are not well represented in the product pool. Furthermore, small fragments below about 200 bp, which would otherwise preferentially amplify, are rarely amplified because of sequence complementarity in the adapters, which promotes the formation of panhandle structures that are stable enough below this size range to greatly reduce amplification. Therefore, it is important to have restriction fragments between 0.2 and 1.2 kb.

We compared the effectiveness of four different restriction enzymes to identify strain-specific ORFs in *H. pylori* strain 26695. Previous studies have relied on a single enzyme in SSH experiments. For the experiments reported here, enzymes were chosen that provide average size fragments of between 200 and 500 bp, as determined by agarose gel electrophoresis (data not shown). Fig. 1 shows the

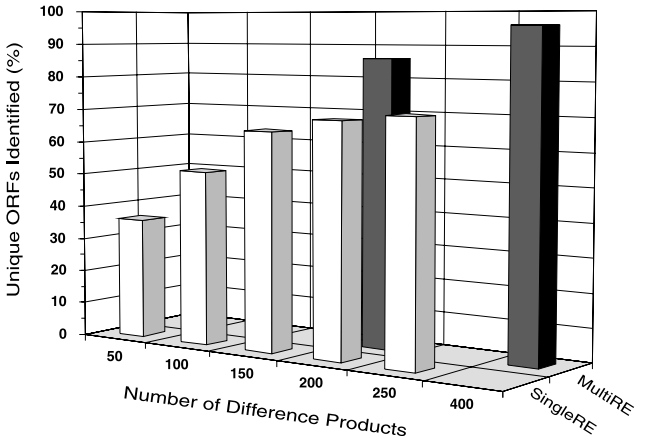

Fig. 2. Comparison of the coverage of unique ORFs by a given number of difference products generated using either a single restriction enzyme (*Alu*I; white columns) or the same number generated using four different restriction enzymes (*Alu*I, *Apo*I, *Dra*I and *Sau*3AI; black columns).

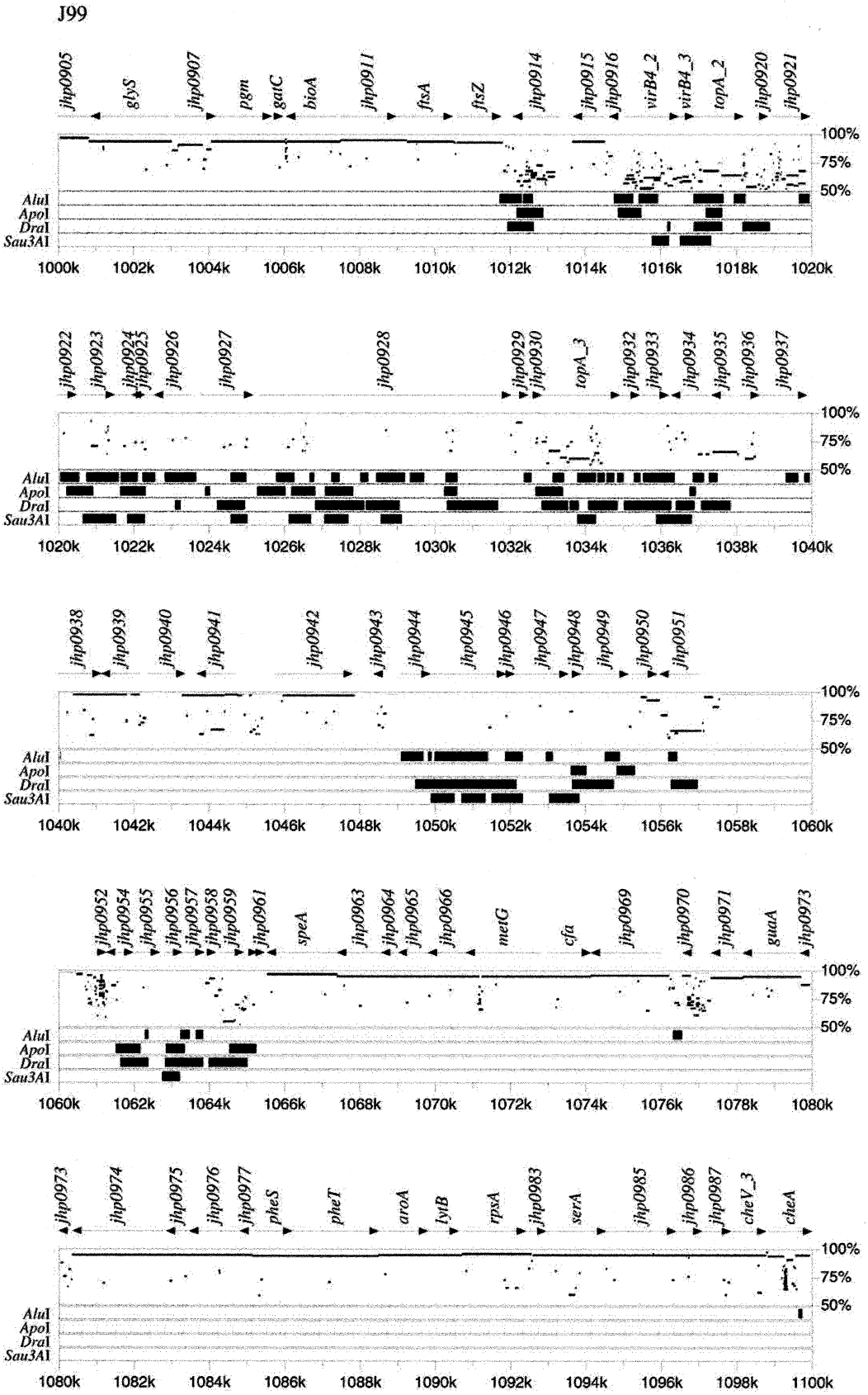

Fig. 3. Percent identity plot (PIP) showing similarity to *H. pylori* 26695 within a 100-kb region of *H. pylori* J99. The tester genome, J99, is plotted horizontally, with ORFs indicated by horizontal arrows. Coordinates in kilobase pairs (k) are shown. Below the tester map the degree of similarity (percent identity) to any region(s) in the driver genome (26695) is indicated by markings on the vertical axis. The black boxes in the appropriate horizontal lane indicate the location of the difference products obtained with each of the restriction enzymes used. PIPs of the entire *H. pylori* genomes are available at http://bio.cse.psu.edu/.

identification of strain-specific ORFs in *H. pylori* 26695 as a function of increasing numbers of *Alu*I difference products, i.e. sequences that showed a normalized cross_match score of 0–0.2. Panel A shows that the curve begins to plateau as more products represent ORFs that have already been identified in that particular experiment. For 150 difference products, 65% of the target ORFs are identified, and this number increases to 70% with 250 products. Panel B shows that the relationship is similar for all enzymes used, *Alu*I, *Apo*I, *Dra*I and *Sau*3AI; 150 difference products for each identify 65% of the strain-specific ORFs. These results suggest that enzymes that generate fragments of the appropriate size show approximately the same degree of coverage.

3.3. Multiple enzymes increase coverage of unique ORFs

Having shown that difference products from the four enzymes used showed similar degrees of coverage in SSH experiments, we next considered whether the use of multiple restriction enzymes would influence coverage when compared to the same number of difference products generated using a single restriction enzyme. We reasoned that using multiple restriction enzymes would increase coverage because in any given difference region the occurrence of some recognition sequences may not produce fragments within the amplifiable size range. Fifty difference products were randomly chosen from each of the four subtractions (using *Alu*I, *Dra*I, *Apo*I and *Sau*3AI) and the combined coverage was compared with 200 *Alu*I difference products (from Fig. 1). In each case *H. pylori* 26695 was the tester strain. Fig. 1 shows that the 200 *Alu*I products identified 69% of the tester-specific ORFs, but 200 products from the four independent subtractions identified substantially more – 86% (Fig. 2). When 400 difference products (100 from each of the four independent subtractions) were analyzed, all but five strain-specific ORFs were identified from the 110 *H. pylori*-specific ORFs (95%) (Fig. 2). These results show that the same number of difference products can yield substantially greater coverage of unique regions when multiple restriction enzymes are used in independent subtraction experiments. Fig. 3 shows the coverage obtained by the multiple enzyme approach in a 100-kb region of the *H. pylori* J99 chromosome using a percent identity plot (PIP) generated by PipMaker [23] to compare the two sequences. The high coverage obtained in both large and small difference regions can be seen, as well as one strain-specific ORF that was not identified (jhp0940). The sequence of jhp0940 indicates that there is one *Alu*I fragment and two *Apo*I fragments within the optimum amplifiable size range that could have generated difference products. This suggests that jhp0940 remained unidentified due to the statistics of genome coverage, and not because of the frequency or location of the restriction enzyme recognition sequences present in this region.

4. Discussion

Prokaryotic genome sequencing requires great effort and cost. It is now appreciated that genomic variation within closely related groups of prokaryotes can be substantial, suggesting a need to define and study these differences, for example in the case of *H. pylori* [24]. We have shown that high throughput subtractive hybridization is an approach that allows comprehensive genomic surveys of strains by directing sequencing to regions that differ among strains. Correlation between phenotypic differences and gene differences is clear in many cases, such as in pathogenic versus non-pathogenic strains. This information will provide a better view of diversity within a closely related group, as well as allowing careful choices for more extensive genome projects including studies directed at particular regions. The difference products can be used for complete sequencing of novel regions when used as hybridization probes to screen libraries, or the sequence information can be used for the amplification and sequencing of flanking sequences [25,26]. Also, such surveys will increase the likelihood of finding species- or strain-specific regions useful for diagnostics.

Genome variation is often associated with the acquisition and deletion of large (10–50-kb) regions of DNA [14,27]. SSH relies on the isolation of restriction fragments, which are contained within such regions. The present study exploited the complete sequences of two *H. pylori* strains to test the ability of subtractive hybridization to survey the genetic differences over a whole genome comprehensively. In scenarios where the tester genome is uncharacterized, it will be straightforward to record the number of times that identical or overlapping clones are sequenced. These data can be used to generate a curve, similar to that of Fig. 1, which will indicate near-complete coverage as the number of previously unidentified subtraction products diminishes. Coverage can be monitored in this fashion, and sequencing proceeds to a degree chosen by the experimenters according to their requirements.

Subtractive hybridization will not provide information about point mutations or genes that are intact in one genome and interrupted in another, but as stated above novel gene content is a crucial aspect of genetic variation in prokaryotes and the main impetus for genome projects. Another limitation of all subtractive hybridization methods is that there will always be a certain proportion of false positives. In the case of SSH, false positives arise from sequences that, in spite of the presence of excess driver containing complementary sequences, remain single-stranded and are able to eventually hybridize with a complementary strand from tester DNA to form an amplifiable product. Identifying false positives requires semi-quantitative hybridization experiments [3,28] or evaluation by PCR amplification [5,6], but these steps can be eliminated when a complete sequence of the driver genome is

available, greatly improving the power of the approach. In the case of SSH, a method has been reported that allows the reduction of false positives, which could reduce the amount of sequencing necessary for genomic surveys [29].

Another valuable and immediate benefit of strain-specific sequences would be to augment information contained within microarrays, thus greatly expanding the scope of either gene content or gene expression analyses. As more information is accumulated, it should be possible to develop broad-based multi-strain chips that would be a much more powerful tool than the single-strain microarrays now in use or under development. One more near-term benefit would be the development of strain-specific DNA-based diagnostic tools for rapid strain detection and identification. In the longer term, strain-specific surface structures could be identified that would provide attractive targets for rapid antibody-based identification assays, and better knowledge of a common core set of genes within a species could help in the development of new antibiotics and vaccines. Strain variation also provides valuable insights into evolutionary processes, and finding sequences that are common among strains will facilitate more precise and reliable taxonomy. Finally, studies of novel genes may help elucidate the basic biology of interesting strain differences, leading to a more fundamental understanding of microbial diversity.

Acknowledgements

We thank Tim McDaniel and Richard Alm for the two *H. pylori* strains used in this study. We would also like to acknowledge the technical expertise of Jessica Wollard, Anne Marie Erler, and Warren Regala. This work was supported by the U.S. Department of Energy's Microbial Genome Program. W.M. was supported by Grant HG02238 from the National Human Genome Research Institute. The portion of this work carried out at Lawrence Livermore National Laboratory was performed under the auspices of the U.S. Department of Energy by the University of California, under Contract W-7405-Eng-48.

References

[1] DeShazer, D., Waag, D.M., Fritz, D.L. and Woods, D.E. (2001) Identification of a *Burkholderia mallei* polysaccharide gene cluster by subtractive hybridization and demonstration that the encoded capsule is an essential virulence determinant. Microb. Pathog. 30, 253–269.

[2] Tinsley, C.R. and Nassif, X. (1996) Analysis of the genetic differences between *Neisseria meningitidis* and *Neisseria gonorrhoeae*: two closely related bacteria expressing two different pathogenicities. Proc. Natl. Acad. Sci. USA 93, 11109–11114.

[3] Emmerth, M., Goebel, W., Miller, S.I. and Hueck, C.J. (1999) Genomic subtraction identifies *Salmonella typhimurium* prophages, F-related plasmid sequences, and a novel fimbrial operon, *stf*, which are absent in *Salmonella typhi*. J. Bacteriol. 181, 5652–5661.

[4] Lin, L.F., Posfai, J., Roberts, R.J. and Kong, H. (2001) Comparative genomics of the restriction-modification systems in *Helicobacter pylori*. Proc. Natl. Acad. Sci. USA 98, 2740–2745.

[5] Agron, P.G., Walker, R.L., Kinde, H., Sawyer, S.J., Hayes da, W.C., Wollard, J. and Andersen, G.L. (2001) Identification by subtractive hybridization of sequences specific for *Salmonella enterica* serovar Enteritidis. Appl. Environ. Microbiol. 67, 4984–4991.

[6] Radnedge, L., Gamez-Chin, S., McCready, P.M., Worsham, P.L. and Andersen, G.L. (2001) Identification of nucleotide sequences for the specific and rapid detection of *Yersinia pestis*. Appl. Environ. Microbiol. 67, 3759–3762.

[7] Townsend, K.M., Frost, A.J., Lee, C.W., Papadimitriou, J.M. and Dawkins, H.J. (1998) Development of PCR assays for species- and type-specific identification of *Pasteurella multocida* isolates. J. Clin. Microbiol. 36, 1096–1100.

[8] Sawada, K., Kokeguchi, S., Hongyo, H., Sawada, S., Miyamoto, M., Maeda, H., Nishimura, F., Takashiba, S. and Murayama, Y. (1999) Identification by subtractive hybridization of a novel insertion sequence specific for virulent strains of *Porphyromonas gingivalis*. Infect. Immun. 67, 5621–5625.

[9] Kato-Maeda, M., Rhee, J.T., Gingeras, T.R., Salamon, H., Drenkow, J., Smittipat, N. and Small, P.M. (2001) Comparing genomes within the species *Mycobacterium tuberculosis*. Genome Res. 11, 547–554.

[10] Behr, M.A., Wilson, M.A., Gill, W.P., Salamon, H., Schoolnik, G.K., Rane, S. and Small, P.M. (1999) Comparative genomics of BCG vaccines by whole-genome DNA microarray. Science 284, 1520–1523.

[11] Murray, A.E., Lies, D., Li, G., Nealson, K., Zhou, J. and Tiedje, J.M. (2001) DNA/DNA hybridization to microarrays reveals gene-specific differences between closely related microbial genomes. Proc. Natl. Acad. Sci. USA 98, 9853–9858.

[12] Salama, N., Guillemin, K., McDaniel, T.K., Sherlock, G., Tompkins, L. and Falkow, S. (2000) A whole-genome microarray reveals genetic diversity among *Helicobacter pylori* strains. Proc. Natl. Acad. Sci. USA 97, 14668–14673.

[13] Blattner, F.R., Plunkett III, G., Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F., Gregor, J., Davis, N.W., Kirkpatrick, H.A., Goeden, M.A., Rose, D.J., Mau, B. and Shao, Y. (1997) The complete genome sequence of *Escherichia coli* K-12. Science 277, 1453–1474.

[14] Perna, N.T. et al. (2001) Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. Nature 409, 529–533.

[15] Alm, R.A., Ling, L.S., Moir, D.T., King, B.L., Brown, E.D., Doig, P.C., Smith, D.R., Noonan, B., Guild, B.C., deJonge, B.L., Carmel, G., Tummino, P.J., Caruso, A., Uria-Nickelsen, M., Mills, D.M., Ives, C., Gibson, R., Merberg, D., Mills, S.D., Jiang, Q., Taylor, D.E., Vovis, G.F. and Trust, T.J. (1999) Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. Nature 397, 176–180.

[16] Tomb, J.F., White, O., Kerlavage, A.R., Clayton, R.A., Sutton, G.G., Fleischmann, R.D., Ketchum, K.A., Klenk, H.P., Gill, S., Dougherty, B.A., Nelson, K., Quackenbush, J., Zhou, L., Kirkness, E.F., Peterson, S., Loftus, B., Richardson, D., Dodson, R., Khalak, H.G., Glodek, A., McKenney, K., Fitzegerald, L.M., Lee, N., Adams, M.D., Hickey, E.K., Berg, D.E., Gocayne, J.D., Utterback, T.R., Peterson, J.D., Kelley, J.M., Cotton, M.D., Weidman, J.M., Fujii, C., Bowman, C., Watthey, L., Wallin, E., Hayes, W.S., Borodovsky, M., Karp, P.D., Smith, H.O., Fraser, C.M. and Venter, J.C. (1997) The complete genome sequence of the gastric pathogen *Helicobacter pylori*. Nature 388, 539–547.

[17] Lisitsyn, N. and Wigler, M. (1993) Cloning the differences between two complex genomes. Science 259, 946–951.

[18] Akopyants, N.S., Fradkov, A., Diatchenko, L., Hill, J.E., Siebert, P.D., Lukyanov, S.A., Sverdlov, E.D. and Berg, D.E. (1998) PCR-

based subtractive hybridization and differences in gene content among strains of *Helicobacter pylori*. Proc. Natl. Acad. Sci. USA 95, 13108–13113.

[19] Straus, D. and Ausubel, F.M. (1990) Genomic subtraction for cloning DNA corresponding to deletion mutations. Proc. Natl. Acad. Sci. USA 87, 1889–1893.

[20] Diatchenko, L., Lau, Y.F., Campbell, A.P., Chenchik, A., Moqadam, F., Huang, B., Lukyanov, S., Lukyanov, K., Gurskaya, N., Sverdlov, E.D. and Siebert, P.D. (1996) Suppression subtractive hybridization: a method for generating differentially regulated or tissue-specific cDNA probes and libraries. Proc. Natl. Acad. Sci. USA 93, 6025–6030.

[21] Skowronski, E.W., Armstrong, N., Andersen, G., Macht, M. and McCready, P.M. (2000) Magnetic microplate-format plasmid isolation protocol for high-yield, sequencing grade DNA. BioTechniques 29, 786–790.

[22] Gordon, D., Abajian, C. and Green, P. (1998) Consed: a graphical tool for sequence finishing. Genome Res. 8, 195–202.

[23] Schwartz, S., Zhang, Z., Frazer, K.A., Smit, A., Riemer, C., Bouck, J., Gibbs, R., Hardison, R. and Miller, W. (2000) PipMaker – a web server for aligning two genomic DNA sequences. Genome Res. 10, 577–586.

[24] Suerbaum, S. and Achtman, M. (1999) Evolution of *Helicobacter pylori*: the role of recombination. Trends Microbiol. 7, 182–184.

[25] Sarkar, G., Turner, R.T. and Bolander, M.E. (1993) Restriction-site PCR: a direct method of unknown sequence retrieval adjacent to a known locus by using universal primers. PCR Methods Appl. 2, 318–322.

[26] Arnold, C. and Hodgson, I.J. (1991) Vectorette PCR: a novel approach to genomic walking. PCR Methods Appl. 1, 39–42.

[27] Lawrence, J.G. and Ochman, H. (1997) Amelioration of bacterial genomes: rates of change and exchange. J. Mol. Evol. 44, 383–397.

[28] Bogush, M.L., Velikodvorskaya, T.V., Lebedev, Y.B., Nikolaev, L.G., Lukyanov, S.A., Fradkov, A.F., Pliyev, B.K., Boichenko, M.N., Usatova, G.N., Vorobiev, A.A., Andersen, G.L. and Sverdlov, E.D. (1999) Identification and localization of differences between *Escherichia coli* and *Salmonella typhimurium* genomes by suppressive subtractive hybridization. Mol. Gen. Genet. 262, 721–729.

[29] Rebrikov, D.V., Britanova, O.V., Gurskaya, N.G., Lukyanov, K.A., Tarabykin, V.S. and Lukyanov, S.A. (2000) Mirror orientation selection (MOS): a method for eliminating false positive clones from libraries generated by suppression subtractive hybridization. Nucleic Acids Res. 28, e90.