NotI passporting to identify species composition of complex microbial systems

Veronika Zabarovska¹, Alexey S. Kutsenko^{1,2}, Lev Petrenko¹, Gelena Kilosanidze^{1,3}, Olle Ljungqvist⁴, Elisabeth Norin¹, Tore Midtvedt¹, Gösta Winberg¹, Roland Möllby¹, Vladimir I. Kashuba¹, Ingemar Ernberg¹ and Eugene R. Zabarovsky^{1,3,*}

¹Microbiology and Tumor Biology Center, Center for Genomics and Bioinformatics, and Department of Cell and Molecular Biology, Karolinska Institute, 171 77 Stockholm, Sweden, ²Center for Theoretical Problems of Physicochemical Pharmacology, Russian Academy of Sciences, 177 977 Moscow, Russia, ³Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, 119 991 Moscow, Russia and ⁴Center of Gastrointestinal Disease, Ersta Hospital, 116 91 Stockholm, Sweden

Received August 9, 2002; Revised and Accepted November 13, 2002

ABSTRACT

We describe here a new method for large-scale scanning of microbial genomes on a quantitative and qualitative basis. To achieve this aim we propose to create *NotI* passports: databases containing *NotI* tags. We demonstrated that these tags comprising 19 bp of sequence information could be successfully generated using DNA isolated from intestinal or fecal samples. Such *NotI* passports allow the discrimination between closely related bacterial species and even strains. This procedure for generating restriction site tagged sequences (RSTS) is called passporting and can be adapted to any other rare cutting restriction enzyme. A comparison of 1312 tags from available sequenced *Escherichia coli* genomes, generated with the *NotI*, *PmeI* and *SbfI* restriction enzymes, revealed only 219 tags that were not unique. None of these tags matched human or rodent sequences. Therefore the approach allows analysis of complex microbial mixtures such as in human gut and identification with high accuracy of a particular bacterial strain on a quantitative and qualitative basis.

INTRODUCTION

There is still much to learn about the human normal microflora. The human gut contains ~1–2 kg of bacterial cells. The number of these cells in the intestine is 10–100 times larger than the number of cells in the human body but at best 10–15% of the microbial species are known, presumably because nobody can grow them *in vitro* in media due to the unknown growth requirements, oxygen sensitivity and other unidentified components (1). Many intestinal bacteria are known to provide molecules that the host itself cannot

manufacture or degrade from nutritional compounds. Thus, these organisms are clearly of survival value, they are true symbionts of the host (2–5).

During the past 10 years enormous progress has been achieved in sequencing of bacterial genomes (6). Different approaches and techniques have been developed and now several bacterial genomes are being sequenced each month. However, despite impressive technological progress, mapping and sequencing of even small bacterial genomes is still expensive and laborious.

After completion of the genomic sequence from representative organisms, there is now a great demand for comparison with the genomes of other individuals, related species, etc. This growing field of comparative genomics is highly relevant for our understanding of human and animal health and disease, epidemiology, evolution and ecology (7–10). An area where this comparative technique can be rapidly and effectively applied is in the sequencing of related bacterial strains and species, in order to identify the genomic basis for differences in their biological properties, particularly pathogenicity, e.g. in the challenging task to study the human intestinal flora and to identify pathogenicity islands (9–12). The current strategies for mapping and sequencing are clearly not meeting the challenge of high-throughput comparative genomics. Alternative and complementary strategies need to be developed, and it is imperative now to find cost-effective and convenient methods that allow comparative genomics projects to be undertaken by a wide range of laboratories.

Recently we have developed a new efficient strategy for simultaneous genome mapping and sequencing. The approach is based on physically oriented, overlapping restriction fragment libraries called slalom libraries (13). Our model experiments have demonstrated the feasibility of the approach, and showed that the efficiency (cost-effectiveness and speed) of existing mapping/sequencing methods could be improved at least 5- to 10-fold. The benefits of the slalom approach will be most obvious in comparative sequencing experiments in

*To whom correspondence should be addressed at Microbiology and Tumor Biology Center, Karolinska Institute, Box 280, 171 77, Stockholm, Sweden. Tel: +46 8 728 6750; Fax: +46 8 319 470; Email: eugzab@ki.se

combination with new high-throughput sequencing technologies which cannot be used in the shotgun sequencing approach. Even the short sequences (e.g. 20 bp) generated by pyrosequencing or massively parallel signature sequencing (MPSS) (14,15) should suffice, in principle. The combination of such novel sequencing techniques with the slalom approach would increase the power of the method 10- to 50-fold more. However, while the slalom approach can be used for the analysis of a particular bacterial species it cannot be used for analysis of complex microbial mixtures.

Identification of bacterial species and strains relies heavily on culture techniques. However, in a complex bacterial population, rapidly growing bacteria would overgrow, making quantification and identification of slow- or non-growing bacteria impossible. Techniques still have to be developed even to culture a representative selection of the microorganisms. Consequently, the picture of the intestinal flora has been biased in favor of the more easily cultured bacteria. There are some other methods available for biochemical analysis of complex microbial mixtures, e.g. by enzyme analysis which requires growth of colonies outside the body, or analysis of the composition of long fatty acids in stools which gives a crude indication of the composition of the normal flora. The limitations of such techniques are obvious (1,3,16).

The most popular approach that has so far been used to classify constituent strains in microbial mixtures without growing them *in vitro* is PCR amplification of the 16S rRNA genes. The amplified material can be used for fingerprinting or/and cloning and sequencing. However, both approaches are rather unspecific: different bacterial species can have the same sequence or PCR fragment and different PCR fragments/sequences can represent the same bacterial species. As 16S rRNA genes are conserved through evolution this approach cannot discriminate between closely related bacterial species or strains (12,16–19). Microarray-based techniques represent another alternative but they also have limitations (20–22). We have shown that the frequency profile of specific short oligonucleotides can be used to correctly classify bacterial genomes with an accuracy of 85%. However, at least 400 bases are needed for this approach (23).

Several directions of our previous and proposed studies were associated with human *NotI* jumping and linking clones (24–26). *NotI* linking clones contain DNA fragments flanking a single *NotI* recognition site, while *NotI* jumping clones contain sequences adjacent to neighboring sites. We have generated more than 22 000 unique human *NotI* flanking sequences (27). Analyses of these sequences have demonstrated that even short sequences surrounding *NotI* sites can yield important information allowing efficient isolation of new genes.

These results led to the realization that it would be feasible to develop a procedure for the analysis of complex microbial mixtures based on the short sequences surrounding *NotI* sites or, more general, on restriction site tagged sequences (RSTS).

MATERIALS AND METHODS

Growth of bacteria, other microbiology and molecular procedures were performed according to standard methods. Genomic DNA extraction from bacterial cells was performed according to the protocol at the [mmi/bbhelp/bac1.html. The protocol was based on the method of Marmur \(1961\) modified by J. L. Johnson \(Virginia Tech, Blacksburg, VA\).](http://www.uct.ac.za/depts/</p></div><div data-bbox=)

The *NotI*-passporting procedure

Two oligonucleotides, BfocII: 5'-GGATGAAAAGTGGGA-3' and Z98NOT: 3'-GTCGTGACTGGGAAAACCCTGGCCT-ACCTTTTGACCTCCGG-5' were used to create the *NotI* linker.

Bacterial DNA (2 µg) at a concentration of 50 µg/ml was digested with 20 U *NotI* (Roche Molecular Biochemicals, Indianapolis, IN) at 37°C for 2 h, and then heat-inactivated for 20 min at 85°C.

Then, 0.4 µg of the digested DNA was ligated to *NotI* linker (50 M excess) overnight with T4 DNA ligase (Roche Molecular Biochemicals) in the appropriate buffer in 100 µl reaction mixtures. The DNA was then concentrated with ethanol and digested with 10 U *BpmI* at 37°C for 3 h.

Following digestion, *BpmI* was heat inactivated and the DNA was ligated overnight in the presence of a 50 M excess of the ZNBpm linker at room temperature. Two nucleotides: Zamine: 5'-CTCAAACCGT-3' amine and Z2_univer: 3'-NNGAGTTTGGCACAGCACTGACCCTTTTGGGACC-5' were used to create ZNBpm linker.

The sample was then purified using a JETquick PCR Purification Spin Kit (Genomed GmbH, Bad Oeynhausen, Germany), and dissolved in 100 µl TE. An aliquot of 1 µl of this sample was PCR amplified with Z1 univer (3'-GAG-TTTGGCACAGCACTGACCCTTTTGGGACC-5') and anti-univer (5'-CAGCACTGACCCTTTTGGGACC-3') primers.

PCR was performed in 40 µl solution containing 67 mM Tris-HCl (pH 9.1), 16.6 mM (NH₄)₂SO₄, 2.0 mM MgCl₂, 0.1% Tween-20, 200 µM dNTPs, 3 µl PCR pool, 400 nM of each primer, and 5 U *Taq* DNA polymerase. The PCR cycling conditions were 95°C for 1.5 min, followed by 25 cycles of 95°C for 20 s, 66°C for 10 s, with 72°C for 10 s, with a final extension at 72°C for 1 min.

The final product was purified with the JETquick PCR Purification Spin Kit (Genomed) and cloned using TOPO TA Cloning kit (Invitrogen, Carlsbad, CA). Sequencing gels were run on ABI 377 automated sequencers (Applied Biosystems, Foster City, CA), according to the manufacturers' protocols, using standard primers.

Sequence analysis

To facilitate RSTS analysis, restriction sites for all commercially available 8-bp restriction enzymes with adjacent 11 bp were extracted from the completed bacterial genome sequences available in EMBL database (release 71). In this way database comprising 19-nucleotide tags was prepared for subsequent analysis.

The analysis of sequences was performed at the Karolinska Institute Sequence Analysis Center (kisac.cgb.ki.se), using local versions of programs and public databases.

The database of nucleotide tags was compared with the draft human genome sequence and EMBL database (release 72) including all human and bacterial entries. Nucleotide similarity searches were performed with BLAST 2.2 (28,29) using the non-gapped alignment method (blast parameter *g* = F). The report cut-off (blast parameter *-b*) for the high-

Figure 1. General scheme of the experiment. Two alternative approaches for the *NotI* passporing are shown: single tag sequencing and concatemer strategy. Two DNA samples are digested with *NotI* and ligated to a linker with *NotI* sticky ends. This linker contains the *BpmI* recognition sites. This restriction nuclease cut 16/14 bp outside of the recognition site. Then ligation mixtures are digested with this enzyme to generate 11/9 nucleotide tags adjacent to the *NotI* site. After several biochemical operations these tags are PCR amplified and ligated into concatemers that are cloned in plasmid vectors and sequenced (central part of the scheme). Alternatively, in single tag approach (shown at the left and the right) these PCR-amplified tags are directly cloned and sequenced. *NotI* tags from both DNA samples are collected and compared to identify the differences.

scoring segment pairs was set to 50. The statistical significance threshold (blast parameter -e) was set to: $e = 1.E-1$.

RESULTS AND DISCUSSION

Rationale for the experiments

The main idea was to develop new efficient methodologies for analysis of complex microbiological systems. As described above for the slalom approach, even short 20 bp sequence tags distributed over the whole genome suffice to generate a minimal contig of overlapping clones and therefore uniquely describe a bacterial genome.

We suggest use of only some specific fragments of the genomes (e.g. *NotI* linking clones, *NotI* tags, etc., refs 24,30) for analysis of the species composition in the intestinal microflora. Thus we do not aim to sequence complete

genomes or study completely sequenced genes. We assign special signatures for the particular microorganism/genes and analyze these signatures in different samples of colonic flora. In this work we analyze the use of short sequence tags adjacent to *NotI* or other restriction enzyme recognition sites. The collection of *NotI* tags represents the *NotI* sequence passport or, in short, *NotI* passport. By the term '*NotI* passporing' we mean the process of creating *NotI* tags/passports.

The naming is based on the enzymes originally used, but the methods can be adapted to other restriction enzymes as well.

The general design of the experiment is as follows (Fig. 1). DNA from fecal samples and surgical specimens is digested with *NotI* and subsequently a *NotI* passport for the particular specimen is generated. Comparing such passports from different individuals or from the same individual before and after drug treatment will reveal the differences between them. This information can in some cases be directly used to make

Figure 2. Flow chart diagram explaining the generation of *NotI* tags. For producing ditags or concatemers the final PCR products may be digested with *FokI* or *FokI* + *BpmI* and ligated. *BpmI* recognition site is shown red and *FokI* in violet. The *NotI* tag sequence in the final PCR product stands out against blue background.

conclusions. In other cases, using these sequences we can identify *NotI* linking clones that are different between two samples. These clones can be used for further analysis e.g. finding the genes that are involved in the examined matter (e.g. effects of antibiotic treatment, chronic bowel disease, etc.) or sequencing/isolation of the required microorganism.

NotI passporting

We originally considered using the serial analysis of gene expression (SAGE) technique for this purpose. SAGE allows for both a representative and comprehensive differential gene expression profile (31). The idea of this approach is that, for each of the mRNA molecules, a short 10 bp sequence tag is produced (including recognition site for the tagging enzyme it is 14 bp) (32). Then these tags are ligated into concatemers and cloned. One sequencing reaction produces information for tens of RNA molecules. Thus by the sequencing of a few thousand clones one can expect to sample e.g. all of the estimated 10 000 to 50 000 expressed genes in a given cell population. Experiments with the SAGE technique for producing *NotI* tags were, however, unsuccessful. The complexity of genomic DNA in microbial mixtures is at least 100 times higher than the complexity of mRNA in eukaryotic cells.

All RNA molecules must be tagged in SAGE but in our case, approximately one out of 250 molecules should be tagged. We propose to produce one tag for each 100–1000 kb, but in SAGE one tag is produced for every 256 bp. At the same time, a 14 bp tag does not suffice for unambiguous identification of sequences in genomic DNA. That is why we have developed a new procedure called *Not* passporting (see Materials and Methods). In this work we used the following modification (Fig. 2). Genomic DNA was digested with *NotI* and ligated to a linker with *NotI* sticky ends. This linker contained the *BpmI* recognition sites. This restriction nuclease cut 16/14 bp outside of the recognition site. The ligation mixture was digested with this enzyme to generate 11/9 nucleotide tags adjacent to the *NotI* site. This DNA sample was ligated to ZNBpm linker and PCR amplified with antiuniver and Z1univer primers to generate a 85 bp duplex. The final PCR-amplified molecule contained a 17 bp sequence tag which was missing 2 bp from the original *NotI* site (Fig. 2) and therefore the whole *NotI* tag contained 19 bp. *NotI* passports were experimentally produced for *Escherichia coli* K12, *Enterobacter cloacae* R4 and *Klebsiella pneumoniae* B4958. Experiments with samples obtained from mice demonstrated that the quality of DNA isolated from intestine or feces was

Table 1. The number of recognition sites for rare cutting restriction enzymes in selected bacterial species

No.	Microbial genomes	Genome length (Mb)	Number of recognition sites						
			<i>NotI</i>	<i>PacI</i>	<i>PmeI</i>	<i>SbfI</i>	<i>SgfI</i>	<i>SgrAI</i>	<i>Sse2321</i>
1	<i>Aeropyrum pernix</i> K1	1.7	88	8	29	62	5	135	48
2	<i>Aquifex aeolicus</i>	1.6	4	37	36	34	2	29	6
3	<i>Bacillus subtilis</i>	4.2	81	89	89	51	51	157	52
4	<i>Borrelia burgdorferi</i>	1.5	1	234	37	8	0	2	0
5	<i>Buchnera</i> sp. APS	0.7	0	184	12	5	2	1	0
6	<i>Campylobacter jejuni</i>	1.6	0	91	42	13	5	1	0
7	<i>Chlamydomophila pneumoniae</i> AR39	1.2	2	59	10	21	13	4	1
8	<i>Clostridium perfringens</i> 13	3.1	1	479	51	6	0	2	0
9	<i>Deinococcus radiodurans</i> R1	3.3	15	1	4	28	7	645	164
10	<i>Escherichia coli</i> K12	4.6	23	143	87	68	222	548	31
11	<i>Escherichia coli</i> O157:H7	5.5	36	165	92	108	239	642	34
12	<i>Helicobacter pylori</i> 26695	1.7	7	32	35	4	88	61	12
13	<i>Helicobacter pylori</i> J99	1.6	14	34	43	4	87	66	15
14	<i>Lactococcus lactis</i> subsp. <i>lactis</i>	2.4	3	176	47	17	2	11	0
15	<i>Mesorhizobium loti</i>	7.6	1434	3	6	395	1041	4829	2686
16	<i>Methanococcus jannaschii</i>	1.7	0	258	40	8	0	7	4
17	<i>Neisseria meningitidis</i> MC58	2.3	74	24	27	28	11	360	176
18	<i>Pasteurella multocida</i> PM70	2.3	5	182	61	12	37	70	3
19	<i>Pyrococcus abyssi</i>	1.8	10	42	20	6	5	32	6
20	<i>Pyrococcus furiosus</i> DSM 3638	1.9	6	63	18	21	0	7	0
21	<i>Rickettsia prowazekii</i>	1.1	1	239	20	10	1	4	0
22	<i>Salmonella typhi</i>	4.8	89	129	79	132	276	768	238
23	<i>Staphylococcus aureus</i> Mu50	2.9	0	440	83	12	5	12	2
24	<i>Streptococcus pneumoniae</i> R6	2.0	1	40	25	30	1	9	0
25	<i>Sulfolobus solfataricus</i>	3.0	3	71	34	11	7	16	0
26	<i>Synechocystis</i> PCC6803	3.6	44	192	104	40	3160	182	18
27	<i>Thermotoga maritima</i>	1.9	10	4	5	14	33	117	7
28	<i>Treponema pallidum</i>	1.1	20	7	11	67	18	167	37
29	<i>Vibrio cholerae</i>	4.0	73	103	117	37	203	199	24
30	<i>Yersinia pestis</i>	4.7	65	207	87	50	111	447	22

Figure 3. Analysis of specificity of RSTS for *E.coli* strains. Ditags and tags unique for *E.coli* species (A). Altogether 656 recognition sites (ditags) produce 1312 tags and among them 219 are non-unique. Analysis of strain-specific tags for *E.coli* K12 (B). Fifty-seven strain-specific restriction sites give 97 unique tags absent in other *E.coli* strains.

Table 2. Specificity of *NotI*, *PmeI* and *SbfI* recognition sites in 70 microbial genomes

No.	Microbial genomes	Genome length (Mb)	Number of recognition sites				Species-specific unique restriction ditags (%)
			<i>NotI</i>	<i>PmeI</i>	<i>SbfI</i>	<i>NotI</i> + <i>PmeI</i> + <i>SbfI</i>	
1	<i>Aeropyrum pernix</i> K1	1.7	88	29	62	179	100.0
2	<i>Agrobacterium tumefaciens</i> Cereon	5.7	609	9	258	876	96.9
3	<i>Agrobacterium tumefaciens</i> Dupont	5.7	608	9	258	875	96.9
4	<i>Aquifex aeolicus</i>	1.6	4	36	34	74	100.0
5	<i>Archaeoglobus fulgidus</i>	2.2	2	10	81	93	100.0
6	<i>Bacillus halodurans</i> C-125	4.2	23	109	22	154	100.0
7	<i>Bacillus subtilis</i>	4.2	81	89	51	221	98.2
8	<i>Borrelia burgdorferi</i>	1.5	1	37	8	46	97.8
9	<i>Brucella melitensis</i> strain 16M	3.3	144	7	53	204	87.7
10	<i>Buchnera</i> sp. APS	0.7	0	12	5	17	88.2
11	<i>Campylobacter jejuni</i>	1.6	0	42	13	55	100.0
12	<i>Caulobacter crescentus</i>	4.0	884	3	331	1218	99.6
13	<i>Chlamydia muridarum</i>	1.1	2	12	13	27	100.0
14	<i>Chlamydia pneumoniae</i> CWL029	1.2	2	10	21	33	100.0
15	<i>Chlamydia trachomatis</i>	1.0	3	13	18	34	100.0
16	<i>Chlamydophila pneumoniae</i> AR39	1.2	2	10	21	33	100.0
17	<i>Chlamydophila pneumoniae</i> J138	1.2	2	10	21	33	100.0
18	<i>Clostridium acetobutylicum</i>	4.1	0	62	16	78	100.0
19	<i>Clostridium perfringens</i> 13	3.1	1	51	6	58	100.0
20	<i>Deinococcus radiodurans</i> R1	3.3	15	4	28	47	95.7
21	<i>Escherichia coli</i> K12	4.6	23	87	68	178	88.2
22	<i>Escherichia coli</i> O157:H7	5.5	36	92	108	236	88.1
23	<i>Escherichia coli</i> O157:H7 EDL933	5.5	36	93	113	242	86.8
24	<i>Haemophilus influenzae</i> Rd	1.8	1	55	15	71	100.0
25	<i>Halobacterium</i> sp. NRC-1	2.6	791	6	81	878	92.8
26	<i>Helicobacter pylori</i> 26695	1.7	7	35	4	46	100.0
27	<i>Helicobacter pylori</i> J99	1.6	14	43	4	61	100.0
28	<i>Lactococcus lactis</i> subsp lactis	2.4	3	47	17	67	100.0
29	<i>Listeria innocua</i> Clip11262	3.0	8	69	2	79	84.8
30	<i>Listeria monocytogenes</i> strain EGD	2.9	8	52	8	68	83.8
31	<i>Mesorhizobium loti</i>	7.6	1434	6	395	1835	99.1
32	<i>Methanobacterium thermoautotrophicum</i> delta H	1.8	7	6	96	109	99.1
33	<i>Methanococcus jannaschii</i>	1.7	0	40	8	48	100.0
34	<i>Mycobacterium leprae</i> strain TN	3.3	251	9	83	343	98.5
35	<i>Mycobacterium tuberculosis</i> CDC1551	4.4	1214	2	163	1379	93.2
36	<i>Mycobacterium tuberculosis</i> H37Rv	4.4	1216	2	164	1382	93.4
37	<i>Mycoplasma genitalium</i> G37	0.6	0	57	0	57	100.0
38	<i>Mycoplasma pneumoniae</i>	0.8	2	62	1	65	98.5
39	<i>Mycoplasma pulmonis</i>	1.0	0	21	0	21	100.0
40	<i>Neisseria meningitidis</i> MC58	2.3	74	27	28	129	90.7
41	<i>Neisseria meningitidis</i> Z2491	2.2	76	26	30	132	92.4
42	<i>Nostoc</i> sp. PCC 7120	7.2	21	129	0	150	96.7
43	<i>Pasteurella multocida</i> PM70	2.3	5	61	12	78	100.0
44	<i>Pseudomonas aeruginosa</i> PA01	6.3	979	1	831	1811	99.5
45	<i>Pyrobaculum aerophilum</i> strain IM2	2.2	111	84	16	211	100.0
46	<i>Pyrococcus abyssi</i>	1.8	10	20	6	36	94.4
47	<i>Pyrococcus furiosus</i> DSM 3638	1.9	6	18	21	45	95.6
48	<i>Pyrococcus horikoshii</i>	1.7	11	23	9	43	95.3
49	<i>Ralstonia solanacearum</i> GMI1000	5.8	1375	1	572	1948	99.7
50	<i>Rickettsia conorii</i> Malish 7	1.3	2	35	11	48	93.8
51	<i>Rickettsia prowazekii</i>	1.1	1	20	10	31	90.3
52	<i>Salmonella typhi</i>	4.8	89	79	132	300	91.0
53	<i>Salmonella typhimurium</i> LT2	4.9	95	83	123	301	94.8
54	<i>Sinorhizobium meliloti</i>	3.7	760	3	215	978	90.3
55	<i>Staphylococcus aureus</i> Mu50	2.9	0	83	12	95	98.9
56	<i>Staphylococcus aureus</i> N315	2.8	0	81	13	94	100.0
57	<i>Streptococcus pneumoniae</i> R6	2.0	1	25	30	56	98.2
58	<i>Streptococcus pneumoniae</i> TIGR4	2.2	1	26	31	58	91.4
59	<i>Streptococcus pyogenes</i> M1 GAS	1.9	3	34	11	48	97.9
60	<i>Sulfolobus solfataricus</i>	3.0	3	34	11	48	100.0
61	<i>Sulfolobus tokodaii</i>	2.7	2	73	12	87	100.0
62	<i>Synechocystis</i> PCC6803	3.6	44	104	40	188	94.1
63	<i>Thermoplasma acidophilum</i>	1.6	9	3	43	55	100.0
64	<i>Thermoplasma volcanium</i>	1.6	3	20	25	48	100.0
65	<i>Thermotoga maritima</i>	1.9	10	5	14	29	100.0
66	<i>Treponema pallidum</i>	1.1	20	11	67	98	100.0

Table 2. Continued

No.	Microbial genomes	Genome length (Mb)	Number of recognition sites				Species-specific unique restriction ditags (%)
			<i>NotI</i>	<i>PmeI</i>	<i>SbfI</i>	<i>NotI</i> + <i>PmeI</i> + <i>SbfI</i>	
67	<i>Ureaplasma urealyticum</i>	0.8	0	25	1	26	100.0
68	<i>Vibrio cholerae</i>	5.1	73	117	37	227	98.2
69	<i>Xylella fastidiosa</i>	2.7	87	22	32	141	99.3
70	<i>Yersinia pestis</i>	4.7	65	87	50	202	91.1
	Total	199.7	11458	2708	5095	19261	96.6

sufficient to obtain *NotI* tags. The *NotI* passports uniquely identified these species and among 96 tags none was common for these three bacterial species. Of course, ditags or concatemers (31) also can be created from these 85 bp products after digestion with *FokI* or *FokI* + *BpmI*. We believe that new high-throughput technologies like MPSS will make sequencing of single tags a more efficient approach than creation of concatemers.

Restriction site tags for the analysis of complex microbial mixtures

As was pointed out above, this restriction site tagging procedure can be adapted to any recognition site for a restriction nuclease. It seems likely that for comprehensive analysis of flora composition use of several passports will be advantageous: different bacteria possess very different CG content. For *NotI* passports this means that bacteria having high CG content (*NotI* recognition site: GCGGCCGC) will be preferentially represented while using for example *SwaI* passports (*SwaI*: ATTTAAAT) will be better for bacterial genomes with high AT content. The choice of passporting strategy may also depend on the bacterial strains expected in connection with studies focused on particular problems such as cancer risk, medication, diet, etc.

We decided to test the potential of the passporting approach and analyzed 70 bacterial species that were completely sequenced (EMBL database, release 71). As an example, Table 1 shows the number of recognition sites for seven rare-cutting restriction enzymes in 30 bacterial species. It is easy to see that all 30 microbial species have different numbers of *NotI* recognition sites and therefore can be distinguished by *NotI* passporting. Moreover, from Table 1 we can see that the *PmeI* (GTTTAAAC) and *SbfI* (CCTGCAGG) restriction enzymes may be even more informative. Table 1 shows that different strains of *E.coli* and *Helicobacter pylori* have various numbers of *NotI*, *PmeI* and *SbfI* recognition sites. Thus all of these strains are uniquely defined by any of these enzymes and therefore this approach can really discriminate between different species and strains, something that is not possible with 16S rRNA genes sequencing (see also Fig. 3).

To further investigate the specificity of *NotI*, *PmeI* and *SbfI* tags, we compared these tags from three completely sequenced *E.coli* strains with all microbial and human sequences (EMBL database, release 72). The three *E.coli* strains contained altogether 1312 tags for these three enzymes and among them only 219 were not unique (Fig. 3A). The tag

was considered unique if it matched only *E.coli* sequences. Importantly, since two tags describe the same restriction (e.g. *NotI*) site, one tag may be unique while the other can be non-unique and therefore both tags (unique and non-unique) may still represent a unique restriction site. In this particular case only 81 from 656 restriction sites were not unique. Figure 3B shows the results of a similar experiment but in this case only *E.coli* K12 tags were used for comparison with all other microbial sequences. Here the tag was considered unique if it matched only *E.coli* K12 sequences. Of course, the number (and proportion) of strain-specific tags was significantly less than the number (and proportion) of species-specific tags (97 unique versus 259 non-unique in Figure 3B against 1093 unique versus 219 non-unique in Figure 3A). In fact, all three *E.coli* strains had strain-specific tags. Even the O157H7 and O157H7 EDL933 strains could be distinguished and each had two unique tags.

These results demonstrate the power of the approach. In our comparative experiments we also analyzed EST and EMBL entries including all sequences. A vast majority of *NotI* tags were unique even if 1–2 nucleotide sequence mismatches were allowed.

As mentioned above a strong feature of *NotI* passporting is the internal controls. If a *NotI* site from a particular bacterial species yields, for example, *NotI* tag 100 and *NotI* tag 101 then we shall obtain both tags in approximately equimolar quantities, i.e. in similar numbers. If the situation is different (only *NotI* tag 100 is present) then it most probably means that *NotI* tag 100 originates from another (novel) bacterial species.

Table 2 shows the fraction of species-specific recognition sites (ditags) for *NotI*, *PmeI* and *SbfI* for 70 completely sequenced bacteria available in October 2002. We considered ditag as species specific if it matched sequences only from the same bacterial species. On average 96.6% of ditags was species specific and *Listeria monocytogenes* strain EGD had the lowest fraction of species-specific tags: 83.8%.

To estimate efficacy of commercially available 8-bp cutters we measured the information value of these restriction enzymes for all 70 sequenced microbial genomes (EMBL database, release 72). If a particular genome contained no recognition sites for the analyzed restriction enzyme then this restriction enzyme got –0.5 points, if it contained 1–2 sites then this added 0 points, for 3–4 sites 0.5 points was added and so on, as follows: 5–105 sites: 1.0 point; 106–200 sites: 0.5 points; 201–500 sites: 0 points; more than 500 sites: –0.5 points. The results for seven enzymes in Table 1 are

Figure 4. Schematic distribution of recognition sites for seven restriction enzymes in 70 completely sequenced bacterial genomes. Differently colored sections indicate different informative values based on the number of recognition sites in a particular genome (see text).

shown in Figure 4 and summary results in Table 3. As it was apparent from Table 1 and Figure 4, *NotI*, *PmeI* and *SbfI* are among the most informative enzymes.

Recently, a new DNA polymerase, *Phi29* became commercially available (Amersham Biosciences, Buckinghamshire, UK). This DNA polymerase can be used for isothermal

Table 3. Information value of recognition sites for rare cutting restriction enzymes in 70 microbial genomes

Restriction enzyme	<i>PmeI</i>	<i>SbfI</i>	<i>PacI</i>	<i>FspA</i>	<i>NotI</i>	<i>SgfI</i>	<i>SgrAI</i>	<i>SrfI</i>	<i>Sse232I</i>	<i>AscI</i>	<i>FseI</i>	<i>SwaI</i>
Score	62.5	49.5	37	35.5	22.5	22.5	21.5	16.5	16	14	13	-21

amplification of large DNA molecules including whole bacterial genomes (33,34). Sequence information obtained during passporting can be used for the design of PCR primers and for amplification of genomes from unknown bacterial species showing interesting features. This means that the RSTS approach gives a possibility both for studying biodiversity and for isolation of new microorganisms. It is also possible to use another enzyme instead of *BpmI* to obtain longer tags (G. Winberg and E. Zabarovsky, unpublished results).

The ability to analyze complex microbial mixtures is of great importance for many applications. For instance, differences between individual compositions of the normal flora will be instrumental for future analysis of the effects on the normal flora composition of diet, foods, geographical location, medication such as antibiotics and use of probiotics. Conversely, effects of intestinal microorganisms on colonic diseases, autoimmunity and colonic cancer risk can be evaluated. We suggest for this type of analysis to use the present approach, generation of RSTS. Hundreds of thousands of tags can be produced by a small research group in a short time allowing careful analysis of thousands of bacterial species/strains (31). We have demonstrated that such *NotI* tags can be efficiently produced and we have proven their high specificity. It is worthwhile to note that we have recently developed a new subtraction method (cloning of deleted sequences: CODE) and shown that this method can be efficiently applied to the *NotI* flanking sequences (30). Thus, the power and sensitivity of the passporting procedure can be significantly increased by removing the most abundant species with the CODE technique (35,36). We suggest to create a database for '*NotI* passports' (as mentioned above, it is may be more correct to speak about 'RSTS passports'). Such a database can be used together with a *NotI* (RST) microarray database (30) as these approaches are mutually complementary. This integrated database most probably will generate new knowledge as these two approaches are based on completely different biochemical techniques but aim to solve the same problem. As bacterial species are very different in nucleotide content we propose using combinations of different types of representation probes/microarrays (*NotI*, *PmeI*) for different applications.

ACKNOWLEDGEMENTS

This work was supported by research grants from the Swedish Cancer Society, the Swedish Research Council, The Swedish Foundation for International Cooperation in Research and Higher Education (STINT), Pharmacia Corporation and Karolinska Institute.

REFERENCES

- Langendijk,P.S., Schut,F., Jansen,G.J., Raangs,G.C., Kamphuis,G.R., Wilkinson,M.H. and Welling,G.W. (1995) Quantitative fluorescence *in situ* hybridization of *Bifidobacterium* spp. with genus-specific 16S rRNA-targeted probes and its application in fecal samples. *Appl. Environ. Microbiol.*, **61**, 3069–3075.
- Finegold,S.M., Sutter,V.L. and Mathisen,G.E. (1983) Normal indigenous flora. In Hentges,D.J. (ed.), *Human Intestine Microflora in Health and Disease*. New York, Academic Press, pp. 3–31.
- Katouli,M., Bark,T., Ljungqvist,O., Svenberg,T. and Mollby,R. (1994) Composition and diversity of intestinal coliform flora influence bacterial translocation in rats after hemorrhagic stress. *Infect. Immun.*, **62**, 4768–4774.
- Gibson,G.R. and Roberfroid,M.B. (1995) Dietary modulation of the human colonic microbiota: introducing the concept of prebiotics. *J. Nutr.*, **125**, 1401–1412.
- Midtvedt,T. (1999) In L.Å. Hansson and R.H. Yolken (eds), *Microbial Functional Activities*. Lippincott-Raven, Philadelphia, Vol. Nestle Nutritional Workshop Series, pp. 79–96.
- Broder,S. and Venter,J.C. (2000) Sequencing the entire genomes of free-living organisms: the foundation of pharmacology in the new millennium. *Annu. Rev. Pharmacol. Toxicol.*, **40**, 97–132.
- Myers,E.W., Sutton,G.G., Delcher,A.L., Dew,I.M., Fasulo,D.P., Flanagan,M.J., Kravitz,S.A., Mobarry,C.M., Reinert,K.H., Remington,K.A. *et al.* (2000) A whole-genome assembly of *Drosophila*. *Science*, **28**, 2196–2204.
- Fenchel,T. (2002) Microbial behavior in a heterogeneous world. *Science*, **296**, 1068–1071.
- Read,T.D., Salzberg,S.L., Pop,M., Shumway,M., Umayam,L., Jiang,L., Holtzapple,E., Busch,J.D., Smith,K.L., Schupp,J.M. *et al.* (2002) Comparative genome sequencing for discovery of novel polymorphisms in *Bacillus anthracis*. *Science*, **296**, 2028–2033.
- Cummings,C.A. and Relman,D.A. (2002) Genomics and microbiology. Microbial forensics—"cross-examining pathogens". *Science*, **296**, 1976–1979.
- Hooper,L.V., Wong,M.H., Thelin,A., Hansson,L., Falk,P.G. and Gordon,J.I. (2001) Molecular analysis of commensal host-microbial relationships in the intestine. *Science*, **291**, 881–884.
- Swidsinski,A., Ladhoff,A., Pernthaler,A., Swidsinski,S., Loening-Baucke,V., Ortner,M., Weber,J., Hoffmann,U., Schreiber,S., Dietel,M. *et al.* (2002) Mucosal flora in inflammatory bowel disease. *Gastroenterology*, **122**, 44–54.
- Zabarovska,V.I., Gizatullin,R.Z., Al-Amin,A.N., Podowski,R., Protopopov,A.I., Lofdahl,S., Wahlestedt,C., Winberg,G., Kashuba,V.I. *et al.* (2002) A new approach to genome mapping and sequencing: slalom libraries. *Nucleic Acids Res.*, **30**, e6.
- Ronaghi,M., Pettersson,B., Uhlen,M. and Nyren,P. (1998) A sequencing method based on real-time pyrophosphate. *Science*, **281**, 363–365.
- Brenner,S., Johnson,M., Bridgman,J., Golda,G., Lloyd,D.H., Johnson,D., Luo,S., McCurdy,S., Foy,M., Ewan,M. *et al.* (2000) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat. Biotechnol.*, **18**, 630–634.
- Suau,A., Bonnet,R., Sutren,M., Godon,J.J., Gibson,G.R., Collins,M.D. and Dore,J. (1999) Direct analysis of genes encoding 16S rRNA from complex communities reveals many novel molecular species within the human gut. *Appl. Environ. Microbiol.*, **65**, 4799–4807.
- Pettersson,B., Fellstrom,C., Andersson,A., Uhlen,M., Gunnarsson,A. and Johansson,K.E. (1996) The phylogeny of intestinal porcine spirochetes (*Serpulina* species) based on sequence analysis of the 16S rRNA gene. *J. Bacteriol.*, **178**, 4189–4199.
- Kraaz,W., Pettersson,B., Thunberg,U., Engstrand,L. and Fellstrom,C. (2000) *Brachyspira aalborgi* infection diagnosed by culture and 16S

- ribosomal DNA sequencing using human colonic biopsy specimens. *J. Clin. Microbiol.*, **38**, 3555–3560.
19. Satokari, R.M., Vaughan, E.E., Akkermans, A.D., Saarela, M. and de Vos, W.M. (2001) Bifidobacterial diversity in human feces detected by genus-specific PCR and denaturing gradient gel electrophoresis. *Appl. Environ. Microbiol.*, **67**, 504–513.
 20. Troesch, A., Nguyen, H., Miyada, C.G., Desvarenne, S., Gingeras, T.R., Kaplan, P.M., Cros, P. and Mabilat, C. (1999) *Mycobacterium* species identification and rifampin resistance testing with high-density DNA probe arrays. *J. Clin. Microbiol.*, **37**, 49–55.
 21. Murray, A.E., Lies, D., Li, G., Neilson, 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.
 22. Haffajee, A.D., Smith, C., Torresyap, G., Thompson, M., Guerrero, D. and Socransky, S.S. (2001) Efficacy of manual and powered toothbrushes (II). Effect on microbiological parameters. *J. Clin. Periodontol.*, **28**, 947–954.
 23. Sandberg, R., Winberg, G., Branden, C.-I., Kaske, A., Ernberg, I. and Coster, J. (2001) Capturing whole-genome characteristics in short sequences using a naïve Bayesian classifier. *Genome Res.*, **11**, 1404–1409.
 24. Zabarovsky, E.R., Boldog, F., Thompson, T., Scanlon, D., Winberg, G., Marcsek, Z., Erlandsson, R., Stanbridge, E.J., Klein, G. and Sumegi, J. (1990) Construction of a human chromosome 3 specific *NotI* linking library using a novel cloning procedure. *Nucleic Acids Res.*, **11**, 6319–6324.
 25. Zabarovsky, E.R., Boldog, F., Erlandsson, R., Allikmets, R.L., Kashuba, V.I., Marcsek, Z., Stanbridge, E., Sumegi, J., Klein, G. and Winberg, G. (1991) New strategy for mapping the human genome based on a novel procedure for construction of jumping libraries. *Genomics*, **11**, 1030–1039.
 26. Zabarovsky, E.R., Gizatullin, R., Podowski, R.M., Zabarovska, V.V., Xie, L., Muravenko, O.V., Kozyrev, S., Petrenko, L., Skobeleva, N., Li, J. et al. (2000) *NotI* clones in the analysis of the human genome. *Nucleic Acids Res.*, **28**, 1635–1639.
 27. Kutsenko, A.S., Gizatullin, R., Al-Amin, A.N., Wang, F., Podowski, R.M., Matushkin, Y., Kvasha, S., Gyanchandani, A., Muravenko, O.V., Protopopov, A. et al. (2002) Analysis of *NotI* flanking sequences: a new tool for verification of the human genome. *Nucleic Acids Res.*, **30**, 3163–3170.
 28. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
 29. Gish, W. and States, D.J. (1993) Identification of protein coding regions by database similarity search. *Nature Genet.*, **3**, 266–272.
 30. Li, J., Protopopov, A., Wang, F., Sentchenko, V., Petushkov, V., Vorontsova, O., Petrenko, L., Zabarovska, V., Muravenko, O., Braga, E. et al. (2002) *NotI* subtraction and *NotI*-specific microarrays to detect copy number and methylation changes in whole genomes. *Proc. Natl Acad. Sci. USA*, **99**, 10724–10729.
 31. Velculescu, V.E., Zhang, L., Vogelstein, B. and Kinzler, K.W. (1995) Serial analysis of gene expression. *Science*, **270**, 484–487.
 32. Zhang, L., Zhou, W., Velculescu, V.E., Kern, S.E., Hruban, R.H., Hamilton, S.R., Vogelstein, B. and Kinzler, K.W. (1997) Gene expression profiles in normal and cancer cell. *Science*, **276**, 1268–1272.
 33. Baner, J., Nilsson, M., Mendel-Hartvig, M. and Landegren, U. (1998) Signal amplification of padlock probes by rolling circle replication. *Nucleic Acids Res.*, **26**, 5073–5078.
 34. Dean, F.B., Nelson, J.R., Giesler, T.L. and Lasken, R.S. (2001) Rapid amplification of plasmid and phage DNA using *Phi* 29 DNA polymerase and multiply-primed rolling circle amplification. *Genome Res.*, **11**, 1095–1099.
 35. Li, J., Wang, F., Zabarovska, V., Wahlestedt, C. and Zabarovsky, E.R. (2000) Cloning of polymorphisms (COP): enrichment of polymorphic sequences from complex genomes. *Nucleic Acids Res.*, **28**, e1.
 36. Li, J., Wang, F., Kashuba, V., Wahlestedt, C. and Zabarovsky, E.R. (2001) Cloning of deleted sequences (CODE): A genomic subtraction method for enriching and cloning deleted sequences. *Biotechniques*, **31**, 788–793.