Design of an *Enterobacteriaceae* Pan-Genome Microarray Chip

Oksana Lukjancenko and David W. Ussery

Center for Biological Sequence Analysis, Department of Systems Biology,
The Technical University of Denmark, 2800 Kongens Lyngby, Denmark

Abstract. Microarrays are a common method for evaluating genomic content of bacterial species and comparing unsequenced bacterial genomes. This technology allows for quick scans of characteristic genes and chromosomal regions, and to search for indications of horizontal transfer. A high-density microarray chip has been designed, using 116 *Enterobacteriaceae* genome sequences, taking into account the enteric pan-genome. Probes for the microarray were checked *in silico* and performance of the chip, based on experimental strains from four different genera, demonstrate a relatively high ability to distinguish those strains on genus, species, and pathotype/serovar levels. Additionally, the microarray performed well when investigating which genes were found in a given strain of interest. The *Enterobacteriaceae* pan-genome microarray, based on 116 genomes, provides a valuable tool for determination of the genetic makeup of unknown strains within this bacterial family and can introduce insights into phylogenetic relationships.

Keywords: *Enterobacteriaceae*, Pan-genome, DNA microarray analysis, gene, *Escherichia coli.*

1 Introduction

The risk of dying from disease caused by a bacterial infection is greater than that associated with any other type of disease, including cancer or heart attacks [1, 2]. Epidemic infectious diseases are the most serious causes of mortality and morbidity worldwide, more than all other diseases combined. Infections contribute to significant economic loss in most parts of the world, including first world countries that have high income and developed surveillance and control systems [3, 4]. Every year thousands of people are infected by bacterial pathogens, most of which are transmitted through food [5]. The outcome from food-borne human infections can range from mild self-limiting diarrhea to severe illness that requires hospitalization. In rare cases, food-borne illnesses are even fatal [5, 6]. Enteric bacteria, particularly *Salmonella enterica* subsp. *enterica*, are among the leading food-borne pathogens [6, 7]. In light of this, the detailed and rapid investigation of enteric pathogens is essential in modern epidemiology and clinical diagnostics.

Enterobacteriaceae are pervasive. They are widespread in the environment, existing in water, soil, food, and plants, as well as in the normal intestinal flora of many animals and humans [8-12]. Pathogens within this group have developed a diversity

J.H. Chan, Y.-S. Ong, and S.-B. Cho (Eds.): CSBio 2010, CCIS 115, pp. 165–179, 2010.

of strategies to overcome protective host barriers in order to invade the host, resist innate immune response, multiply in specific and normally sterile body sites, and damage cells in order to establish and maintain a successful infection [13, 14]. Genera within *Enterobacteriaceae* family are of interest, as well, because of problems from food spoilage and for that reason are of considerable economic importance [15].

Bacterial genomes vary in size, even among the strains of the same species. Bacterial species can be characterized by its pan-genome. As defined by Tettelin *et al.*, the microbial pan-genome is a complete collection of various genes located within populations at a particular taxonomic level, commonly within a species. The pan-genome concept can of course be expanded to higher levels, such as genus or even a bacterial family. The pan-genome includes a core-genome, which is a minor fraction of the entire gene pool that is shared between all the given strains. Furthermore, there is a much larger, dispensable portion of bacterial genes, that are missing in one or more strains. Also there are some genes that appear to be unique to each strain [16, 17]. Strain-specific genes can, even among a particular species, make up a notably large portion of the pan-genome [18].

Many methods have been developed for characterizing genetic variation. Use of DNA microarrays is becoming a standard procedure for evaluating genotyping – that is, looking at the genetic content of a bacterial species. The price for microarrays used for genotyping was historically expensive, but now is becoming competitive with the cost of other commonly used typing methods, such as previously widely used multilocus sequence typing (MLST). Moreover, it is becoming increasingly popular, quick, and cost-effective to define the presence and absence of each of the assigned genes in the pan-genome of a species. Thus, microarrays, imprinted with all the genes from species' pan-genome can be used to compare and characterize the genomic content of unknown bacterial isolates and to achieve accurate typing information, that can be useful in epidemiological investigations and clinical diagnostics [1, 19]. For instance, array comparative genomic hybridization (aCGH) is frequently used in human cancer studies to genotype cell lines by determination of gene loss and copy number variations [20] or to detect single nucleotide polymorphisms at target loci [21]. Additionally, microarrays have been widely used in human screenings for the determination and genotyping of bacterial species. Microarrays have changed considerably since they were first introduced. Early microarrays for the *E. coli* genome consisted of long fragments of chromosomal DNA (~1000 to 2000 base-pairs), attached to a microscope slide. Later, Affymetrix made an array covering the entire *E. coli* K-12 genome using a set of 10 to 15 probes (synthetic 25mers) for each gene [22], followed shortly by an array which contained 4 *E. coli* genomes [23, 24]. Custom-designed NimbleGen chips have been made including 7 and then 32 *E. coli* genomes [25, 26].

This study describes the design and use of a high-density oligonucleotide microarray covering the pan-genome of 116 genomes within the *Enterobacteriaceae* family. Probes are designed to distinguish among organisms at the level of genera, species, and even single strains. Moreover, probes for determination of particular gene families, comprising *Enterobacteriaceae* pan-genome, are defined. The performance of this microarray is evaluated both *in silico* and experimentally. Its utility is illustrated for the hybridization of genomic DNA in order to compare uncharacterized isolates which have not been sequenced with the 116 known, sequenced strains. A microarray chip approximating the complete pan-genome of *Enterobacteriaceae*

provides optimal sensitivity to characterize isolates. Gene family microarray analysis is useful for medical and environmental diagnoses and will provide an alternative to costly genome libraries, as well as to the sequencing of environmental samples.

2 Materials and Methods

2.1 Bacterial Strains

In this study, one hundred and twelve complete *Enterobacteriaceae* genome sequences and four in progress, which were publically available in GenBank database at the time of analysis (February, 2010), were used for custom microarray design. An overview of the used strains is shown in Table 1 and the complete collection of the strains is described in supplementary Table S1[1].

Table 1. *Enterobacteriaceae* genera used in the design of the microarray chip

Genus	Number of strains	Genus	Number of strains
Buchnera	6	*Photorhabdus*	2
Citrobacter	3	*Salmonella*	18
Cronobacter	2	*Serratia*	1
Dickeya	3	*Shigella*	8
Edwardsiella	2	*Sodalis*	1
Enterobacter	2	*Wigglesworthia*	1
Escherichia	35	*Xenorhabdus*	1
Klebsiella	4	*Yersinia*	14
Pectobacterium	3	*Erwinia*	4
Proteus	3	*Candidatus**	3

* *Candidatus* is not a genus; however some strains were included as they were classified as *Enterobacteriaceae* at the time of study.

Twelve bacterial strains included in experimental evaluation of the chip are listed in Table 3 (Results section).

2.2 Pan-Genomics

The pan-genome was estimated, as described by Snipen *et al* [27]. Briefly, all protein sequences were compared by BLASTP [28]. Two proteins were attributed to a single gene family if they satisfied the 50/50 rule, meaning that when they could produce a pairwise BLASTP alignment covering at least 50% amino of the length of the longest protein with at least 50% of amino acid identity. Each genome was compared successively: for each n additional genome, that genome was compared to any combinations of n-1 genomes and the number of identical 'core genes' and 'genome specific genes' (specific for genome n) were counted for each n. All cumulative BLASTP hits found in the whole set of genomes were plotted as a running total and were considered as pan-genome, which increases as more genomes are added. The number of gene families with at least one representative in every genome was plotted for the core-genome.

[1] Available at http://www.cbs.dtu.dk/~dave/Supplementary_TableS1.pdf

2.3 The Custom-Microarray Design

The custom probe set for the microarrays was designed around 78 different groups of genomes (the list of groups is presented in the Results section, Table 2) including a collection of generic probes for the entire enteric core (97 genes), as well as for the probes that differentiate each genus within *Enterobacteriaceae*. The custom probe set was followed by more specialized probe sets for species-specific classification within *Klebsiella*, *Salmonella*, *Escherichia*, *Shigella*, and *Yersinia* genera and further probe groups were specific for strain and pathotype for *Escherichia coli* genus. Additionally, sets of probes for all the gene families, comprising pan-genome, were included. The custom microarrays, manufactured by NimbleGen, were based on the NimbleGen 12-plex platform.

2.4 Constructing Target Gene Sets

The genome sequences in this study (Table S1) were searched for genes using the Prodigal gene-finding approach [29] in order to standardize gene finding. All protein-coding sequences were aligned all-against-all using BLASTP [28], and similarity was decided according to 50/50 rule. Proteins that satisfy this rule were assigned to one protein family. 'Group specific gene families' (as described above) were found using batch Perl script, which outputs a list of gene families that are either common to or complementary to the genomes included in pan- and core-genome plots (depending on whether unique or core genes are extracted). Representative sequences from each gene network were selected by choosing the organism from which the genes should be extracted. Unique genes were considered to be those that appeared to be conserved only among the strains belonging to a particular group.

2.5 Probe Selection for Target Genes

Probes for target genes were selected using the OligoWiz program, previously described by Wernersson *et al.* [30][31]. At each position along all the input sequence, the suitability of placing a probe was evaluated according to several criteria: melting temperature (ΔT_m), cross-hybridization, folding (self-annealing), position (within the transcript), and 'low-complexity' (absence of subsequences that occur very commonly in the genome/transcriptome). The weighting scores for these criteria are as follow: cross-hybridization, 39%; ΔT_m, 26%; folding, 13%; position, 13%; and low-complexity, 9%. No probes were accepted unless an overall score of at least 0.3 was obtained, and all probes were required to have a length in the range of 42 bp to 50 bp. OligoWiz was originally designed for single genome use, and thus, the program was modified in order to make the mechanisms screening for cross-hybridization less strict as described by Vejborg *et al.* [32]. A new modified scheme included a log-transformation in the underlying calculations. The net effect is insignificant near the upper boundary of the score, but next to the lower boundary it increases the discriminatory power of the tool.

$$\text{BLAST max score} = 1 - \sum_{n}^{i=1} \log(1 + \sum_{m}^{m=1} \frac{hm,i}{100} \quad (1)$$

2.6 Probe Evaluation *in silico*

Probes were aligned against a database consisting of all possible gene sequences in the total data set using BLASTN. The affinity of each probe for every gene was determined and expressed as the number of identical base pairs and by the E-value. Sequences for which the E-value was lower than 0 were extracted using a batch Perl script. Probes that matched strains not expected to belong to particular group were excluded from the further analysis. If more than ten probes per gene remained available after filtering, only not-overlapping ones were used for subsequent analysis. This resulted in the reduction of candidate probes from 106,657 to 53,644. Consequently, the number of probes targeting each gene ranged from 3 to 14 with a median coverage of about 7 probes per gene.

2.7 DNA Preparation and Hybridization

All the experimental isolates were kindly provided by the laboratory of Frank Møller Aarestrup (DTU Food, The Technical University of Denmark). All test strains were grown overnight on blood agar and genomic DNA was isolated as described in the protocol for the Easy-DNA kit from Invitrogen [33]. The method used is briefly described here: the lysis of the cells was performed by the addition of solution A and subsequent incubation at 65°C. Proteins and lipids were precipitated and extracted by the addition of solution B and chloroform. The solution was then centrifuged to separate the solution into two phases. The DNA was in the upper, clear aqueous phase, the proteins and lipids were in the solid interface, and the chloroform formed the lower phase. The DNA was then removed, precipitated with ethanol, and re-suspended in TE buffer.

The genomic DNA was labeled with cy3 dye and hybridized to NimbleGen custom arrays according to Arrays User's Guide for CGH analysis as provided by the manufacturer of the arrays (Roche NimbleGen, Madison, Wisconsin, USA).

2.8 Analysis Methods

In the initial step, the raw data from multiple microarrays was extracted using NimbleScan software, developed by Roche NimbleGen, and combined as a single input. Data analysis was performed in R (a statistical software program), using the 'oligo' package for analyzing oligonucleotide arrays at the probe level. The package was obtained from Bioconductor [34]. The probes were mapped to each gene group, including position, according to the design. Chip analysis workflow then continued as follows:

1. Performance of probe-level normalization using robust multi-array average (RMA) algorithm. RMA method had a three-step procedure consisting of background correction, normalization, and summarization to obtain gene-level relative intensity measures from probe-level intensities [35].
2. Estimation of gene 'on/off' status based on the summarized gene relative intensities and the median of these intensities for each of the 78 groups.

Supporting microarray chip design information is publicly available[2].

[2] http://www.cbs.dtu.dk/~dave/Microarray_Chip_Design_Lukjancenko_2010.ndf

3 Results

3.1 Pan-Genome and Core-Genome Estimation

For each of the considered bacterial strains listed in Table S1 (Supplementary data), the genome sequence was downloaded from NCBI/GenBank. Genes were predicted by Prodigal [29], and translated into proteins. This resulted in a dataset of 887,184 entries with considerable redundancy due to the presence of the same gene in multiple genomes. To reduce the homology, proteins were grouped into the gene families. Proteins were considered conserved (belonging to the same gene group) if they showed at least 50% amino acid identity in a BLASTP alignment covering at least 50% of the length of the longest protein. The combined pan-genome of 116 genomes within *Enterobacteriaceae* was estimated and appeared to contain 44,838 gene families. The core-genome, that is, the number of conserved genes present in all 116 genomes, was estimated to be comprised of 97 conserved gene families.

3.2 Probe and Microarray Design

In the presented *Enterobacteriaceae* pan-genome microarray design strategy, the probe set was designed around 78 different groups of genomes. The microarray was made up of a collection of probes for each genus within *Enterobacteriaceae,* being species-specific for *Klebsiella*, *Salmonella*, *Escherichia*, *Shigella*, and *Yersinia* genera; strain and pathotype specific for *Escherichia coli* genus; core genes; and all protein families, comprising pan-genome. Using the data from the pan- and core-genome estimation step, the number of 'group-specific' genes and probes was determined and are shown in Table 2. Genes were considered to be 'group-unique' if they were found only within genomes, belonging to a particular group, and were absent in all of the rest genomes among a set of 116 genomes.

The final result was a set of 52,356 *Enterobacteriaceae* target sequences, representing genes of both specific groups and pan-genome gene families. The oligos were then selected using OligoWiz [31] based on several criteria, including their specificity, self-annealing, presence of low-complexity sequences, and their lengths adjusted so as to standardize the hybridization strength. Probes were filtered in order to avoid complimentarity with unwanted targets. In the end a set of 130,540 non-overlapping probes with an average length of 49 bp were obtained. The average number of probes per target gene was about 7, although the actual number for any given target depended on the length of the sequence, since shorter sequences have space for fewer non-overlapping probes. For set of probes that represent gene families an average of 3 probes per family was used.

3.3 Validation of the Custom Arrays

The chip design was evaluated by analyzing and comparing hybridization data from twelve control strains, shown in Table 3. Microarray data can have noise, coming from multiple variations which can occur during the array manufacturing process, the preparation of the biological sample for the hybridization, the hybridization of the samples to the array itself, and the quantification of the spot intensities [35]. To remove such variation, which obviously will affect the measured gene intensity levels,

Table 2. Number of 'group specific' gene families and probes before and after *in silico* validation

Probe group	Number of genes before validation	Number of probes before validation	Number of genes after validation	Number of probes after validation
Buchnera genus	14	200	14	123
Candidatus strains	41	584	41	373
Citrobacter genus	20	171	15	95
Cronobacter genus	271	3224	270	2002
Dickeya genus	155	2129	155	1398
Edwardsiella genus	318	3803	317	2447
Enterobacter genus	40	511	40	318
Erwinia genus	217	2919	217	1840
Escherichia genus	1	15	1	10
Escherichia coli 042	106	1047	79	450
Escherichia coli 536	142	1207	95	436
Escherichia coli 55989	72	646	45	272
Escherichia coli APEC	116	1287	14	83
Escherichia coli APEC O1	116	1287	14	83
Escherichia coli Avirulent	69	508	39	241
Escherichia coli B phylogroup	14	175	14	100
Escherichia coli CFT073	292	2251	115	393
Escherichia coli E24377A	249	1700	90	511
Escherichia coli EAEC	72	646	45	272
Escherichia coli ED1a	159	1545	146	823
Escherichia coli EHEC	21	173	13	27
Escherichia coli EPEC	142	1685	126	893
Escherichia coli ETEC	249	1700	90	511
Escherichia coli ExPEC	52	392	17	131
Escherichia coli HS	90	642	44	313
Escherichia coli IAI1	67	499	39	238
Escherichia coli IAI39	77	609	48	262
Escherichia coli K-12	11	159	11	113
Escherichia coli O103:H2	65	693	50	377
Escherichia coli O111:H-	148	1536	54	250
Escherichia coli O127:H6	142	1685	126	893
Escherichia coli O157:H7	68	709	52	379
Escherichia coli O26:H11	74	690	48	280
Escherichia coli S88	52	392	17	131
Escherichia coli SE11	178	1692	70	360
Escherichia coli SE15	58	609	49	328
Escherichia coli SMS-3-5	145	1064	106	501
Escherichia coli UMN026	113	1026	85	505
Escherichia coli UPEC	121	983	49	179
Escherichia coli UTI89	85	754	35	192
Escherichia/Shigella genera	15	184	15	113
Klebsiella genus	242	3296	242	2090
Klebsiella pneumoniae 342	11	93	8	50
Klebsiella pneumoniae MGH 78578	21	237	14	49
Klebsiella pneumoniae NTUH-K2044	339	2636	233	863

Table 2. (*Continued*)

Klebsiella variicola At-22	115	1282	110	758
Pectobacterium genus	166	2287	166	1422
Proteus genus	355	4782	355	3006
Photorhadbus genus	318	4392	318	2728
Salmonella genus	69	933	69	575
Salmonella enterica Agona	136	1151	111	568
Salmonella arizonae	477	3828	474	2245
Salmonella enterica Choleraesuis	92	804	44	87
Salmonella enterica Dublin	101	526	22	77
Salmonella enterica Enteritidis	20	217	9	55
Salmonella enterica Gallinarum	10	88	5	14
Salmonella enterica Heidelberg	91	608	51	249
Salmonella enterica Newport	189	1967	111	351
Salmonella enterica Paratyphi A	10	80	7	10
Salmonella enterica Paratyphi B	436	1982	175	547
Salmonella enterica Paratyphi C	54	266	20	47
Salmonella enterica Schwarzengrund	139	1025	122	498
Salmonella enterica Typhi	69	759	63	326
Salmonella enterica Typhimurium	9	113	3	30
Serratia genus	780	10393	780	6777
Shigella boydii	19	164	16	52
Shigella dysenteriae	113	1216	98	348
Shigella flexneri	17	218	17	123
Shigella genus	28	401	25	178
Shigella sonnei	48	531	32	152
Sodalis genus	420	5697	420	3464
Wigglesworthia genus	212	3029	212	1789
Xenorhabdus genus	82	855	82	527
Yersinia genus	97	4189	97	809
Yersinia enterocolitica	336	1312	336	2655
Yersinia pestis	7	26	5	5
Yersinia pseudotuberculosis	23	165	13	24
Core genes	97	1378	97	850
Gene families	42151	180219	27536	76896

normalization was performed. A set of twelve arrays (one 12plex array) used in the experiment was printed at the same time, so background noise effects were expected to be reasonably similar across all arrays. Only one out of the twelve the results were not as anticipated. The single exception being for the *Salmonella enterica* serovar Choleraesuis isolate, which shows variation. Thus it was decided to exclude hybridization data of this isolate from further analysis. RMA normalization, performed for microarray data of the remaining eleven samples, made the distribution of probe intensities for each array in a set of arrays nearly the same.

In the workflow of further microarray data analysis, the evaluation of which genus, species, pathotype/serovar or strain, the experimental isolate is most likely to be similar to. For each of the seventy-eight gene sets, the median of signal intensities were calculated. The analysis was performed based on both distribution of probe log intensities and the signal median. The examples are shown in Figures 1-3, which visualize

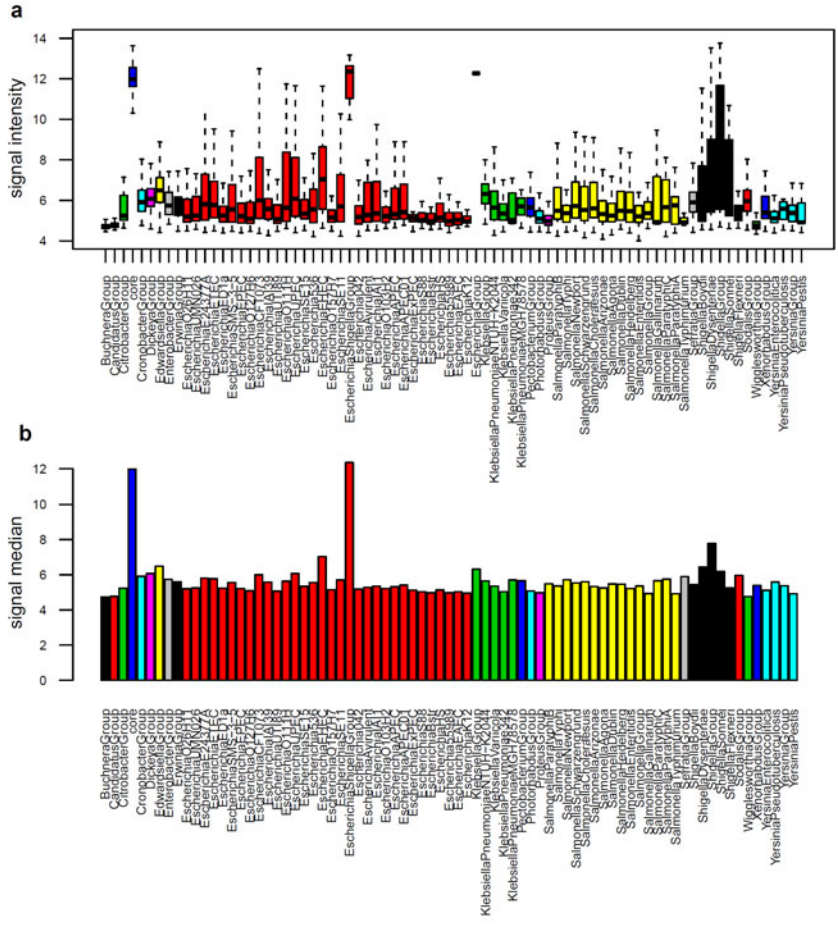

Fig. 1. Distribution of signal intensity and signal median for Escherichia coli ECOR20 strain among the set of seventy-eight groups, mentioned previously in Table 2. a. Box-and-whisker plot, showing signal intensity distribution. b. Bar plot, showing expression signal median distribution. X-axis elements are sorted by genus, based on the order showed in Table 2. Colour code is based on the genera, where 12-colour palette represents 20 genera.

the resulting plots for single representative of three chosen genera *Escherichia*, *Salmonella* and *Yersinia*. Those were *Escherichia coli* ECOR20, *Salmonella enterica* serovar Dublin and *Yersinia frederiksii*, respectively. Table 3 overviews the results for all the eleven isolates, used in the study.

Both box-and-whisker and bar plots for *Escherichia coli* ECOR20, represented in Fig. 1, show high signal intensity among the genes comprising core and *Escherichia*-and-*Shigella* groups. Additionally, results show high similarity to several pathogenic *E. coli* strains, such as *Escherichia coli* CFT073, and strains of O111:H-, UPEC and EHEC pathotypes. Apart from being highly expressed among the genes belonging to *Escherichia* genus, microarray data show relatively high signal level to *Shigella* genus

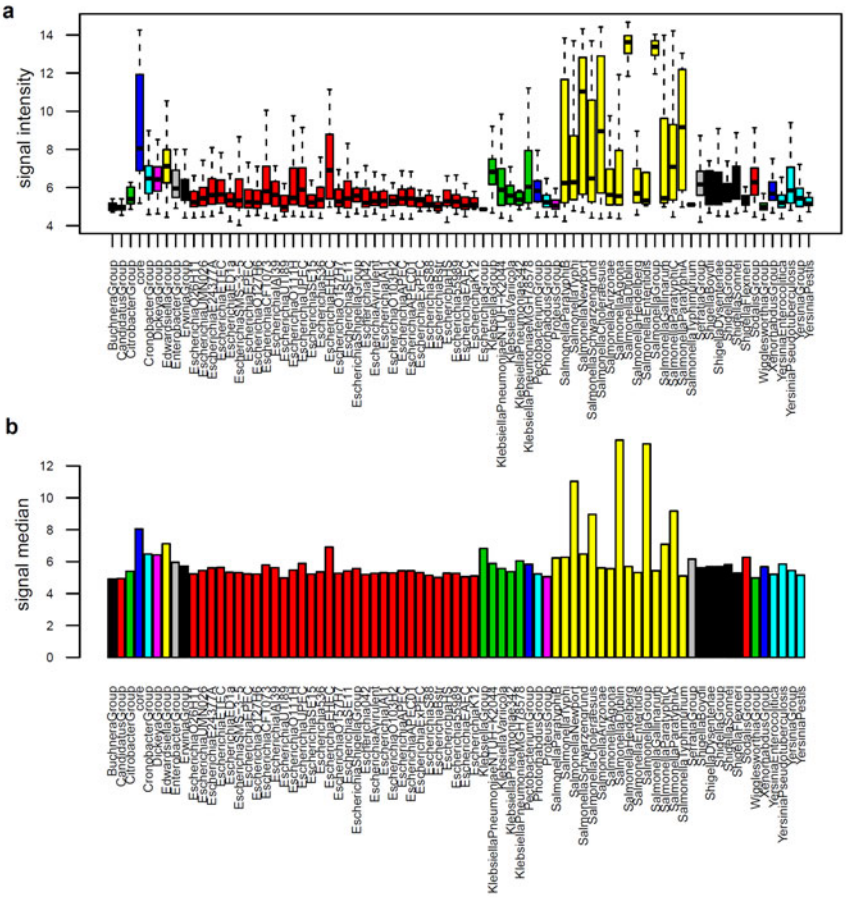

Fig. 2. Distribution of signal intensity and signal median for Salmonella enterica serovar Dublin strain among the set of seventy-eight groups, mentioned previously in Table 2. a. Box-and-whisker plot, showing signal intensity distribution. b. Bar plot, showing expression signal median distribution. X-axis elements are sorted by genus, based on the order showed in Table 2. Colour code is based on the genera, where 12-colour palette represents 20 genera.

strains, thus, resulting in another proof of *Escherichia* and *Shigella* genera strains being very similar.

Fig. 2 visualizes the comparison of data for *Salmonella enterica* serovar Dublin isolate. Genes have high intensity values within strains belonging to *Salmonella* genus and core group. The highest similarity is shown to be Dublin serovar; however, DNA sequences appeared to hybridize with the high strength to Newport, Choleraesuis and Paratyphi A serovar representing probes as well.

In the case of the chosen representative for *Yersinia* genus, *Yersinia frederiksi*, results, shown in Fig. 3, are not that positive, since any obvious high intensity signal cannot be seen. This might occur as a consequence of impropriate isolation of genomic DNA, low concentration of labeled DNA, which was obviously not enough for proper hybridization to target genes, or cross-hybridization effect.

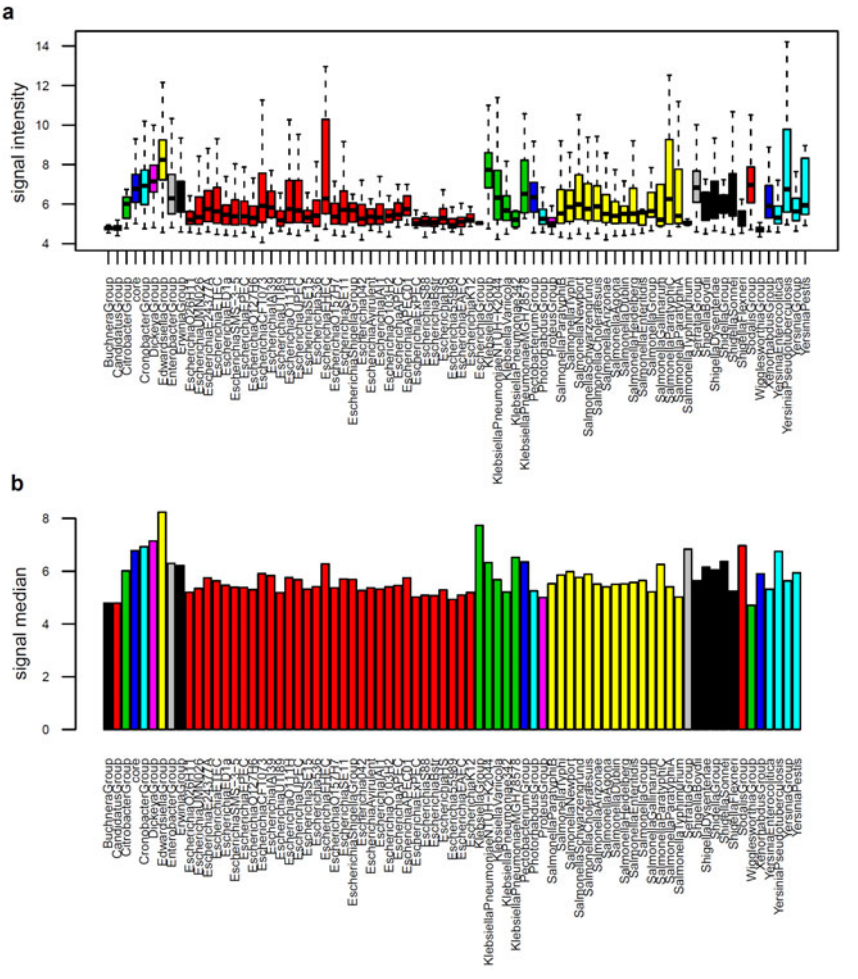

Fig. 3. Distribution of signal intensity and signal median for Yersinia frederiksii strain among the set of seventy-eight groups, mentioned previously in Table 2. a. Box-and-whisker plot, showing signal intensity distribution. b. Bar plot, showing expression signal median distribution. X-axis elements are sorted by genus, based on the order showed in Table 2. Colour code is based on the genera, where 12-colour palette represents 20 genera.

Isolates, results for which are presented in Table 3, show different chip performances. Several of them can be easily proved to belong to a particular genus, specific species and be most likely similar to a particular genus, species or serovar/serotype.

However, some samples, likewise *Yersinia frederiksii*, do not show obvious results. This can consider the presence of uncertainties included in genomic DNA purification and sample preparation for the hybridization.

Table 3. Overview of experimental validation results

Isolate / Distinguishing level	Genera	Species	Pathotype/Serovar
Escherichia coli ECOR20	+	+	-
Salmonella enterica serovar Dublin D6	+	+	+
Salmonella enterica serovar Paratyphi B var Java b	+	+	+
Salmonella enterica serovar Isangi 2005-60-2087-1	+	+	
Salmonella enterica Typhimurium HN-GSS-2007-016	+	+	+
Salmonella enterica serovar Choleraesuis 2870/08			
Shigella sonnei phase 12006-077	-	-	
Shigella flexneri 4 2006-054	+	+	
Shigella boydii 9S	-	-	-
Yersinia entericolitica O3 98-30624-5	-	-	-
Yersinia ruckerii NCTC 10476	-	-	-
Yersinia frederiksii P963	-	-	-

'+' is a positive result, '-' is a negative result and absence of any mark means no analysis with this purpose was made or results are not analysed

4 Discussion and Perspective

The design of a microarray chip covering 116 bacterial genomes has proven to be a considerable challenge. Multiple aspects had to be examined, such as the number of possible sequences to be included in the database, various criteria to select the unique set of genes to particular groups of genomes, and to design probes for them. The greatest difficulty was to optimize these criteria and to filter out the false positive representative sequences for each sequence of interest. Some genera within *Enterobacteriaceae*, such as *Escherichia* and *Shigella*, are quite similar, thus it was difficult to find genus-specific genes. For example, the *Escherichia* genus appeared to have only a single gene family conserved among all the strains belonging to this genus, and being absent in the other enterics. Thus it was an obvious decision to design probes for *Escherichia*-and-*Shigella* genera-specific genes.

Along with choosing representative sequence for each of unique gene family, a problem of selecting the right organism to extract representative sequences for core-genome set became evident. In this study, core-genome genes were extracted from type species of the type genus *Escherichia coli* K-12 MG1655 strain. The unique sets of genes were selected on protein level, that is, similarity/dissimilarity was based on alignment using BLASTP, and gene family members were considered based on the 50/50 rule, described above. Thus this might be an explanation of why some probes did not show high intensity levels at the DNA level as was predicted.

Selecting the probes is indeed a challenging aspect. On the one hand, probes should cover all versions of the same gene, however, at the same time they should be able to distinguish between different genera, species, pathotypes/serovars, and strains. Furthermore, the array should allow various numbers of probes per gene in order to acquire the sufficient coverage of genes. Longer sequences require higher numbers of probes, whereas design of the same number of probes for short genes would result in low quality probes [36]. Therefore, the challenge is to find the best possible solution, with least time, money, and personal energy consumption.

Several improvements and suggestions could be considered for the design of an *Enterobacteriaceae* pan-genome microarray chip. To obtain more sufficient unique gene finding, searchs should be done on DNA level with an appropriate cut-off value. Alignment using the BLASTN algorithm would be able to efficiently identify homologous nucleotide sequences based on similarity and would be helpful in avoiding non-specific probes.

Furthermore, for the validation of the chip step, sample preparations, such as genomic DNA isolation, labeling, and preparation to hybridize an array should be done according to protocols. Purity of DNA should be checked before the DNA labeling step to avoid small quantities of labeled DNA, which hybridizes to wrong sequences and fails to recognize the expected target sequence.

5 Conclusion

In this study, an *Enterobacteriaceae* pan-genome microarray chip was developed based on 116 genomes within this bacterial family. The typical genome size (with the exception of the reduced endosymbiont genomes of *Buchnera*, *Wigglesworthia* and *Sodalis* genera) contained between 3500 and 5500 genes. This made it possible to find at least 10 genus-, species- and pathotype/serovar-genes among all the analysed genomes. This resulted in 53644 unique probes, which were expected to hybridize to particular target sequence. High-density pan-genome microarrays can be very useful in both characterizing DNA content and monitoring expression levels for thousands of genes simultaneously. The comparison of two or more arrays can display the distinct patterns of gene expression or signal intensity level that are useful in the definition of unknown strains or genes included in these genomes. Using some experimental tests the ability of the microarray to determine bacterial strains within *Escherichia* spp., *Shigella* spp., *Salmonella* spp. and *Yersinia* spp. was demonstrated. Most of the results showed discriminative power, although some samples did not show a clear connection to the bacterial strain they are most likely to be similar to. This could be due to low quality DNA from the experiment.

It can be concluded that a *Enterobacteriaceae* pan-genome microarray, based on 116 genomes provides a perfect tool for determination of the genetic makeup of unknown strains within this bacterial family and can introduce insights into phylogenetic relationships.

Acknowledgments. This work is supported by grants from the Danish Center for Scientific Computing and the Danish Research Council. The authors would like to thank Colleen Ussery for help in editing the manuscript.

References

1. Hall, B.G., Ehrlich, G.D., Hu, F.Z.: Pan-genome analysis provides much higher strain typing resolution than multi-locus sequence typing. Microbiology 156, 1060–1068 (2010)
2. Sørensen, T.I., Nielsen, G.G., Andersen, P.K., Teasdale, T.W.: Genetic and environmental influences on premature death in adult adoptees. N. Engl. J. Med. 318, 727–732 (1988)

3. Helms, M., Vastrup, P., Gerner-Smidt, P., Mølbak, K.: Short and long term mortality associated with foodborne bacterial gastrointestinal infections: registry based study. BMJ 326, 357 (2003)
4. Ternhag, A., Törner, A., Svensson, A., Ekdahl, K., Giesecke, J.: Short- and long-term effects of bacterial gastrointestinal infections. Emerging Infect. Dis. 14, 143–148 (2008)
5. Mead, P.S., Slutsker, L., Dietz, V., McCaig, L.F., Bresee, J.S., Shapiro, C., Griffin, P.M., Tauxe, R.V.: Food-related illness and death in the United States. Emerging Infect. Dis. 5, 607–625 (1999)
6. Litrup, E., Torpdahl, M., Malorny, B., Huehn, S., Helms, M., Christensen, H., Nielsen, E.M.: DNA microarray analysis of *Salmonella* serotype Typhimurium strains causing different symptoms of disease. BMC Microbiol. 10, 96 (2010)
7. Laupland, K.B., Schønheyder, H.C., Kennedy, K.J., Lyytikäinen, O., Valiquette, L., Galbraith, J., Collignon, P.: *Salmonella enterica* bacteraemia: a multi-national population-based cohort study. BMC Infect. Dis. 10, 95 (2010)
8. Cheng, S., Hu, Y., Zhang, M., Sun, L.: Analysis of the vaccine potential of a natural avirulent Edwardsiella tarda isolate. Vaccine 28, 2716–2721 (2010)
9. Lindberg, A.M., Ljungh, A., Ahrné, S., Löfdahl, S., Molin, G.: *Enterobacteriaceae* found in high numbers in fish, minced meat and pasteurised milk or cream and the presence of toxin encoding genes. Int. J. Food Microbiol. 39, 11–17 (1998)
10. Musgrove, M.T., Northcutt, J.K., Jones, D.R., Cox, N.A., Harrison, M.A.: *Enterobacteriaceae* and related organisms isolated from shell eggs collected during commercial processing. Poult. Sci. 87, 1211–1218 (2008)
11. Stiles, M.E., Ng, L.K.: *Enterobacteriaceae* associated with meats and meat handling. Appl. Environ. Microbiol. 41, 867–872 (1981)
12. Wright, C., Kominos, S.D., Yee, R.B.: *Enterobacteriaceae* and *Pseudomonas* aeruginosa recovered from vegetable salads. Appl. Environ. Microbiol. 31, 453–454 (1976)
13. Cossart, P., Sansonetti, P.J.: Bacterial invasion: the paradigms of enteroinvasive pathogens. Science 304, 242–248 (2004)
14. Hornef, M.W., Wick, M.J., Rhen, M., Normark, S.: Bacterial strategies for overcoming host innate and adaptive immune responses. Nat. Immunol. 3, 1033–1040 (2002)
15. Olsson, C., Ahrné, S., Pettersson, B., Molin, G.: DNA based classification of food associated *Enterobacteriaceae* previously identified by biology microplates. Syst. Appl. Microbiol. 27, 219–228 (2004)
16. Glasner, J.D., Marquez-Villavicencio, M., Kim, H., Jahn, C.E., Ma, B., Biehl, B.S., Rissman, A.I., Mole, B., Yi, X., Yang, C., Dangl, J.L., Grant, S.R., Perna, N.T., Charkowski, A.O.: Niche-specificity and the variable fraction of the *Pectobacterium* pan-genome. Mol. Plant Microbe Interact 21, 1549–1560 (2008)
17. Tettelin, H., et al.: Genome analysis of multiple pathogenic isolates of *Streptococcus* agalactiae: implications for the microbial ”pan-genome”. Proc. Natl. Acad. Sci. U.S.A. 102, 13950–13955 (2005)
18. Lefébure, T., Stanhope, M.J.: Evolution of the core and pan-genome of *Streptococcus*: positive selection, recombination, and genome composition. Genome Biol. 8, 71 (2007)
19. Phillippy, A.M., Deng, X., Zhang, W., Salzberg, S.L.: Efficient oligonucleotide probe selection for pan-genomic tiling arrays. BMC Bioinformatics 10, 293 (2009)
20. Pinkel, D., Segraves, R., Sudar, D., Clark, S., Poole, I., Kowbel, D., Collins, C., Kuo, W.L., Chen, C., Zhai, Y., Dairkee, S.H., Ljung, B.M., Gray, J.W., Albertson, D.G.: High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. Nat. Genet. 20, 207–211 (1998)

21. Wang, D.G., Fan, J.B., Siao, C.J., Berno, A., Young, P., Sapolsky, R., Ghandour, G., Perkins, N., Winchester, E., Spencer, J., Kruglyak, L., Stein, L., Hsie, L., Topaloglou, T., Hubbell, E., Robinson, E., Mittmann, M., Morris, M.S., Shen, N., Kilburn, D., Rioux, J., Nusbaum, C., Rozen, S., Hudson, T.J., Lipshutz, R., Chee, M., Lander, E.S.: Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science 280, 1077–1082 (1998)
22. Khodursky, A.B., Peter, B.J., Cozzarelli, N.R., Botstein, D., Brown, P.O., Yanofsky, C.: DNA microarray analysis of gene expression in response to physiological and genetic changes that affect tryptophan metabolism in Escherichia coli. Proc. Natl. Acad. Sci. U.S.A. 97, 12170–12175 (2000)
23. Wei, Y., Lee, J.M., Richmond, C., Blattner, F.R., Rafalski, J.A., LaRossa, R.A.: High-density microarray-mediated gene expression profiling of Escherichia coli. J. Bacteriol. 183, 545–556 (2001)
24. Jacobsen, L., Durso, L., Conway, T., Nickerson, K.W.: *Escherichia coli* O157:H7 and other E. coli strains share physiological properties associated with intestinal colonization. Appl. Environ. Microbiol. 75, 4633–4635 (2009)
25. Willenbrock, H., Fridlyand, J.: A comparison study: applying segmentation to array cgh data for downstream analyses. Bioinformatics 21, 4084–4091 (2005)
26. Willenbrock, H., Petersen, A., Sekse, C., Kiil, K., Wasteson, Y., Ussery, D.W.: Design of a seven-genome *Escherichia coli* microarray for comparative genomic profiling. J. Bacteriol. 188, 7713–7721 (2006)
27. Snipen, L., Almøy, T., Ussery, D.W.: Microbial comparative pan-genomics using binomial mixture models. BMC Genomics 10, 385 (2009)
28. Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J.: Gapped BLAST and psi-BLAST: a new generation of protein database search programs. Nucleic. Acids. Res. 25, 3389–3402 (1997)
29. Hyatt, D., Chen, G., Locascio, P.F., Land, M.L., Larimer, F.W., Hauser, L.J.: Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11, 119 (2010)
30. Wernersson, R., Nielsen, H.B.: OligoWiz 2.0–integrating sequence feature annotation into the design of microarray probes. Nucleic Acids Res. 33, W611–W615 (2005)
31. Wernersson, R., Juncker, A.S., Nielsen, H.B.: Probe selection for DNA microarrays using OligoWiz. Nat. Protoc. 2, 2677–2691 (2007)
32. Vejborg, R.M., Bernbom, N., Gram, L., Klemm, P.: Anti-adhesive properties of fish tropomyosins. J. Appl. Microbiol. 105, 141–150 (2008)
33. Easy-DNA kit (2010), `http://tools.invitrogen.com/content/sfs/manuals/easydna_man.pdf`
34. Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A.J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J.Y.H., Zhang, J.: Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 5, R80 (2004)
35. Do, J.H., Choi, D.: Normalization of microarray data: single-labeled and dual-labeled arrays. Mol. Cells 22, 254–261 (2006)
36. Willenbrock, H., Hallin, P.F., Wassenaar, T.M., Ussery, D.W.: Characterization of probiotic *Escherichia coli* isolates with a novel pan-genome microarray. Genome Biol. 8, 267 (2007)