UCRL-JRNL-236055

A functional gene array for detection of bacterial virulence elements

C. Jaing

November 1, 2007

PLoS ONE

Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

A functional gene array for detection of bacterial virulence elements

Crystal Jaing,[1] Shea Gardner,[2] Kevin McLoughlin,[2] Nisha Mulakken,[2] Michelle Alegria-Hartman,[1] Chitra Manohar,[3] Phillip Banda,[1] Peter Williams,[2] Pauline Gu,[2] Mark Wagner,[2] and Tom Slezak[2]

[1]Chemistry, Materials, Earth and Life Sciences, [2]Computation, Lawrence Livermore National Laboratory, Livermore, CA 94550, [3]Celera, Alameda, CA 94502, USA

Corresponding author contact information:
Crystal Jaing
Lawrence Livermore National Lab
Biosciences and Biotechnology
7000 East Ave.
L-371
Livermore, CA 94550
Phone: 925-424-6574
Fax: 925-422-3512
Email: jaing2@llnl.gov

Key word:

Virulence, microarray, NimbleGen, Antibiotic resistance, probe, pathogen

Disclaimer
This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

ABSTRACT

We report our development of the first of a series of microarrays designed to detect pathogens with known mechanisms of virulence and antibiotic resistance. By targeting virulence gene families as well as genes unique to specific biothreat agents, these arrays will provide important data about the pathogenic potential and drug resistance profiles of unknown organisms in environmental samples. To validate our approach, we developed a first generation array targeting genes from *Escherichia coli* strains K12 and CFT073, *Enterococcus faecalis* and *Staphylococcus aureus*. We determined optimal probe design parameters for microorganism detection and discrimination, measured the required target concentration, and assessed tolerance for mismatches between probe and target sequences. Mismatch tolerance is a priority for this application, due to DNA sequence variability among members of gene families. Arrays were created using the NimbleGen Maskless Array Synthesizer at Lawrence Livermore National Laboratory. Purified genomic DNA from combinations of one or more of the four target organisms, pure cultures of four related organisms, and environmental aerosol samples with spiked-in genomic DNA were hybridized to the arrays. Based on the success of this prototype, we plan to design further arrays in this series, with the goal of detecting all known virulence and antibiotic resistance gene families in a greatly expanded set of organisms.

[Supplemental material is available online at www.genome.org]

INTRODUCTION

Rapid detection and characterization of bacterial and viral pathogens has become a vital component of our national biodefense strategy. Various detection technologies based on nucleic acid signatures have emerged in the past few years, including TaqMan and Luminex bead systems. While these technologies are able to rapidly identify selected pathogens at the species or strain level, they do not have the capability to provide broad functional information about known and novel organisms. Characterization of emerging, engineered, or unknown pathogens requires a platform that can assess the virulence and antibiotic resistance mechanisms present in these organisms. One approach that has been used successfully to measure other types of microbial capabilities is known as a functional gene array.

A functional gene array (FGA) is a DNA microarray containing probes targeting sequences unique to genes within families of interest. Family-specific probes are designed to match regions that are conserved among genes in the family, in order to increase the chance of detecting previously unidentified homologs. Small-scale FGAs have been used successfully to measure the presence and activity of key enzymes in environmental samples (Gentry *et al*., 2006). The largest functional gene array described to date contains 1,662 50-mer oligonucleotide probes for 2,402 genes involved in biodegradation and metal resistance (Rhee *et al*., 2004), and was recently upgraded to include over 24,000 probes (He *et al* , 2007). More recently FGAs have been applied in the area of molecular and clinical diagnostics for pathogens (Palacios *et al*, 2007).

The FGAs developed to date have focused on specific sets of gene functions, thereby limiting their use to narrowly defined applications. Constructing a functional gene array to detect genes associated with virulence and antibiotic resistance is a much greater challenge, because of the broad diversity of pathogens and the large number of gene families involved. We loosely define virulence-related genes as those whose products affect the ability of a pathogen to infect or survive in the host, are required for expression of other virulence factors, or cause the host direct harm (such as toxins). A high-density oligonucleotide microarray is the only platform available at present that supports the number of probes required to interrogate such a wide variety of genes. The approach of using presence or absence of virulence genes as a forensic classifier has been demonstrated in a recent study to differentiate several *E. coli* strains by PCR (Jackson et al, 2006). However, the number of genes to screen by PCR is very limited.

Designing and constructing such arrays is typically an expensive and time-consuming process, with some technologies requiring four photolithographic masks for each base position in the probe. This effort would have to be multiplied over several generations of arrays, as we refined our probe design process. Furthermore, the leading provider of high-density oligo arrays (Affymetrix) only supports 25-mer probes, which we determined would not be sensitive enough to detect virulence genes with the required tolerance of mismatches to detect novel homologs. Therefore, the project of constructing an FGA for virulence related genes only became feasible when our laboratory acquired a NimbleGen array synthesizer. The NimbleGen system allows us to do rapid and affordable prototyping of microarrays with up to 385,000 probes per array, with variable probe lengths ranging from 23 to 85 nucleotides (nt).

The NimbleGen arrays are built on glass slides using a proprietary Maskless Array Synthesizer (MAS) system (Singh-Gasson *et al*, 1999). A programmable digital micromirror device projects a pattern of UV light on the slide, which catalyzes a photodeprotection reaction, resulting in the attachment of specific nucleotides to oligos within the illuminated features. The pattern of illumination at each stage of oligo synthesis is controlled by software; therefore no masks need to be constructed, and the cost per array design is greatly reduced. This has made it possible for our group to prototype a series of arrays and to perform experiments with them, in order to find the optimal probe design parameters for detecting signatures of functional gene families.

The other major challenge to be overcome in constructing a functional array for virulence gene detection is the computation required to design sensitive and specific probes for hundreds of thousands of gene target sequences. Millions of sequence comparisons are required to find the most conserved regions within gene families and subfamilies, to ensure that probes are selected to span diverse gene sequences that encode similar functions. Thermodynamic binding energy predictions, conservation and uniqueness scores must be computed for millions of candidate probes, in order to select an optimal combination of probes for each target gene family, balancing sensitivity, specificity, and breadth of coverage. The computation of each of these factors is CPU-intensive, requiring that we develop highly efficient algorithms and implement them using high performance computers (HPC) at LLNL. Our access to HPC facilities has played a crucial role in making high-quality probe design and selection feasible at this scale.

In this report, we describe the process we used to design our first generation functional array for virulence and antibiotic resistance gene families, and the results from our initial experiments with the array. We discuss the probe design algorithms, including virulence gene sequence selection, and our protocols for sample preparation, amplification, labeling, hybridization, and data analysis. We present the results from experiments designed to assess whether the array can detect virulence gene orthologs from organisms without perfect match probes on the array, using both targeted mismatch probes and hybridizations to DNA from other organisms. Also, we report the results from limit of detection studies using bacterial DNA spiked into BioWatch aerosol samples at known concentrations.

RESULTS

Target probe specificity

To assess the ability of the NimbleGen array to reliably identify a target organism of known genome sequence, we performed BLAST searches for all target probe sequences against the four genomes, and selected subsets of probes that had a full length perfect match to one genome, and no perfect match longer than 16 nt to any of the other 3 genomes. We refer to these as strain specific probes. We performed hybridizations with purified DNA from *E. coli* CFT073, *E. faecalis* and *E. coli* K12. The plots in **Figure 1** show $\log_2$ intensities plotted against $\Delta G_{complement}$ for the strain-specific probes for the first two hybridizations. In each case, the probes that were specific to virulence genes present in the target strain had much higher signal intensities than the random control probes and probes specific to the other three organisms. The same pattern was observed in the hybridization with *E. coli* K12 (data not shown). The true positive rate of detection, measured by the fraction of probes specific to the hybridized strain with intensity over the threshold (median + 4 SD) is 100%. The false positive rate – the fraction of intensities over the threshold for probes specific to a different strain – was only 0.29%.

Detection of virulence genes from related organisms

In order to assess the ability of probes designed against gene family members from one organism to detect orthologs with non-identical sequences from other organisms, we performed two sets of experiments, one using the perfect match (PM) target probes, the other (to be discussed below) using the mismatch probes. In the first set of experiments, arrays were hybridized to DNA from four bacterial strains that do not have organism-specific probes on the array, as described in Methods. These four strains were chosen because they have fully sequenced genomes, because they were readily available from ATCC, and because they span a range of phylogenetic distances from the four target strains used to design probes on the array. One of these is a different strain of the same species (*E. coli*) as two of the target strains; one is a different species of the same genus (*Staphylococcus*) as a target strain; one belongs to a different genus (*Salmonella*) of the same family as *E.coli*; and one (*S. pyogenes*) belongs to a different family of the same order as *E.faecalis*. All four strains were found by our HMM analysis to possess orthologs for a variety of virulence gene families; through the results of this analysis, we were able to divide the 299 gene families with probes on the array into groups expected to be present or absent in a given hybridized strain.

The probe intensities are plotted against gene families as shown in **Figure 2**. For clarity, gene families are not labeled in the plots; they are displayed from left to right in arbitrary order. A detailed listing of gene families, their presence or absence and detection or non-detection for each strain is given in **supplementary table S1**. Plot points are colored according to the target genome from which the probe sequence was extracted. The orange line in each plot is the detection threshold for each array (median + 4 SD of random controls).

Figures 2A and **2B** show probes for gene families expected to be present and absent, respectively, in *E. coli* O157:H7 strain EDL933. In this plot, we see intensities above the threshold for 209 of 216 gene families expected to be present in *E. coli* O157:H7, and for 22 out of 83 families lacking orthologs in this strain. Not surprisingly, all of the probes for gene families expected to be present that had positive signals were drawn from one of the two *E. coli* target genomes. We noted only weak signals ($\log_2$ intensity < 11) for 18 of the gene families without orthologs, suggesting that these false positive detection events may be due either to nonspecific hybridization, or to weak cross-hybridization to members of other gene families. BLAST analysis of the probes with strong signals targeting the remaining four families without orthologs showed that the probes matched regions in other genes of *E. coli* O157:H7 encoding domains that are shared between families. For example, several probes for the *fla(A,B,C,E)* family, which is not represented in the O157:H7 EDL933 strain but does contain a flagellin N domain, had matches to the *fliC* gene, which shares this domain.

Figures 2C and **2D** show probes for gene families with and without orthologs, respectively, in *S. saprophyticus* strain ATCC 15305. In this plot, we see intensities above the threshold for 44 of 94 gene families expected to be present, and only 7 of 205 families expected to be absent. Three of these seven have probe intensities exceeding the upper range of the random control probes; of these, two belong to the major facilitator superfamily (MFS) of transporters. Subsequent BLAST analysis revealed that the probes in these families with strong signals matched a 40 base region of a putative *S. saprophyticus* MFS permease gene, with 3 mismatches.

For *S.enterica* serovar *paratyphi*, the majority (126 of 197) of the gene families expected to be present (**Figure 2E)** had probes above the detection threshold; only 7 of the 102 families without orthologs in this strain (**Figure 2F)** had signals above the threshold, and only 3 with strong signals. Again, BLAST analysis of the probes with strong signals showed alignments to other genes with similar functions to those of the targeted gene families.

For *S. pyogenes*, which has the greatest phylogenetic distance from any organism with probes on the array, only 10 of the 53 families with orthologs in this strain (**Figure 2G)** had probes with signals over the threshold; only 3 of these had intensities that exceeded the upper range of the random controls. Probes in these 3 families were drawn from the genomes of both *E. faecalis* and *S. aureus*, which belong to the same class as *S. pyogenes*. Of the 246 families expected to be absent in *S. pyogenes* **(Figure 2H)**, only 5 had probes above threshold, most with low signals ($\log_2$ intensity < 10.5).

Our results from this small set of organisms suggest that probes designed according to our strategy against gene family members from one species can reliably detect orthologs in different species of the same genus, and even different genera of the same taxonomic family. The false positive rate was extremely low in all hybridizations performed.

Mismatch probe sensitivity

To more comprehensively assess the factors influencing the balance between probe sensitivity and specificity, we analyzed data from two series of probes, containing single and multiple mismatches respectively. We first examined data from probes that perfectly matched the hybridized strain, except for a single mismatch (MM) base placed at a known position. We investigated the effect of the mismatch position on the probe intensity, relative to the intensity of the corresponding perfect match (PM) probe for the same hybridization. **Figure 3** shows mean intensity ratios (MM/PM) with error bars corresponding to 95% confidence intervals, averaged over 60 PM probes and their cognate MM probes, from 10 arrays hybridized to a variety of samples. The mismatch positions are numbered in 5’ to 3’ order, and probe lengths range from 30 to 66.

As shown in this figure, single mismatch probe intensities varied with the position of the mismatch. On NimbleGen prokaryotic arrays, the 3’ end of the probe is attached by a 5-T linker to the glass surface of the array. We observed that mismatches located 7 to 20 nt from the 5’ end of the probe had the strongest negative impact on hybridization, while mismatches located on one of the 12 nucleotides closest to the linker had virtually no discernable effect. We also note that, even at the position of maximum effect, 15 nt from the 5’ end, single mismatches have relatively small impact, with only a 35% reduction of intensity relative to the corresponding PM probe.

In the single-mismatch experiments, MM probes were generated containing all three possible choices of mismatch base at each position. We found no consistent difference in intensity between probes using the complement of the PM base and probes generated by transition or non-complementary transversion of the PM base.

When probes contained multiple mismatches to the genome of the hybridized strain, we found that the reduction in intensity depended not only on the number of mismatched bases, but also on the length of the longest perfect match sequence between mismatches. Consequently, longer probes tend to be more tolerant of mismatches. The relationship between the number of mismatch bases, the longest PM region length, and the reduction in intensity relative to the cognate PM probe is shown in **Figure 4**. The graph shows the mean MM/PM intensity ratios averaged over 60 PM probes and the corresponding random mismatch probes. The random mismatch probes were generated from the PM probe sequences by selecting 2, 3, 6, 10, 15 or 20 random positions in the perfect match probe, and creating single mismatches at each position. Intensity ratios were averaged over 10 arrays for the 60 sets of PM and MM probes, and are plotted here on a log scale against the length of the longest PM region.

We observed that probes with two or three mismatches to the hybridized strain had at least half the intensity of the related PM probes, provided there was at least one PM region with length ≥ 29 nt. This was nearly always the case for 60-mer probes. Probes with six mismatches had greater signal reduction, but still had 30% or more of the

PM probe intensity when the mismatches were clustered toward one end of the probe, leaving a 29 nt or longer PM region.

Probes with 10 or more mismatch bases showed even greater signal reduction, and also more variability in reduction between probes. We conjecture that this variability is related to the position of the PM regions within the probe, with regions overlapping the 5’ half of the probe having a stronger positive effect on signal intensity. Additional experiments using a wider variety of mismatch configurations would be required to test this hypothesis adequately.

In general, all probes with PM regions ≥ 29 nt had intensities above the detection threshold. Probes with shorter maximal PM regions were detected some of the time, but not consistently.

Limit of detection of genomic DNA in an environmental sample background

Several experiments were performed to show the dynamic range and limit of detection of our array, along with its ability to identify specific organisms within a complex background, when combined with our protocol for sample preparation. We created six samples using DNA from an aerosol sample (24 hour filter collection from an urban environment) as background material. One sample contained background DNA only; the others were spiked with fragmented *S. aureus* DNA in amounts ranging from 0.31 fg to 3.1 pg, amplified and hybridized to arrays, as described in Methods. For comparison, the complete 2.88 Mb *S. aureus* chromosome has a mass of about 2.95 fg.

Figure 5 shows the intensity of strain-specific probes versus $\Delta G_{complement}$ for arrays hybridized to each of the six samples. In the unspiked aerosol background DNA, we found only a few probes with signals barely above the detection threshold; therefore we expect that the signal seen in the spiked samples is mostly or entirely due to the added *S. aureus* DNA. With 0.31 fg of *S. aureus* DNA, we observe about 36% of *S. aureus* –specific probes with signals above the threshold. The detectable probes cover about 37% of the targeted virulence gene orthologs. With 3.1 fg, we see that 100% of the *S. aureus* specific probes were above the detection threshold. With 31, 310 or 3100 fg, virtually all of the *S. aureus* specific probes were saturated, with intensities within a factor of two below the maximum possible intensity.

DISCUSSION

The emerging threat presented by novel pathogens, whether they arise naturally or are deliberately engineered, creates a need for detection systems that can warn public health authorities about a potential outbreak and help them select appropriate countermeasures. Ideally, such a system will be able to determine the virulence and antibiotic resistance mechanisms present in a sample of unidentified microorganisms, even when the sample includes organisms never previously encountered. As a step toward developing such a system, we have produced and tested a series of highly sensitive and specific functional gene arrays using the NimbleGen platform. These are the first functional gene arrays created that can quantify the presence or absence of hundreds of virulence gene families with a single assay. Our goals for this study were to develop methods for design of gene family-specific probes, to measure the sensitivity and specificity of these arrays, and to assess the validity of the FGA approach for detecting virulence and antibiotic resistance mechanisms in unknown as well as known microorganisms.

The array described in this report includes probes for 1,245 genes, belonging to 299 virulence and antibiotic resistance gene families, identified in *E. coli* K12, *E. coli* CFT073, *S. aureus* and/or *E. faecalis*. These genes were selected using a collection of over 700 hidden Markov models, each of which was trained against sequences of a single virulence or antibiotic resistance gene family, identified by an extensive literature search. Using these models, more than 200,000 gene homologs were identified in a database of bacterial, viral and other genome sequences; the 1,245 genes with probes on the current array are those identified in one of the four target strains. While this set of models targets a substantial fraction of the virulence and antibiotic resistance gene families known at present, future arrays in this series will be based on a comprehensive set of over 1,500 HMMs covering the majority of known virulence and A/R related genes. (McLoughlin, manuscript in preparation).

We used a novel approach to design groups of probes specific for virulence gene families. Prior to our study, there was no software available that could design minimal sets of family-specific probes for such a wide variety of sequences from unrelated organisms. Sequences within a gene family frequently are highly polymorphic at the nucleotide level, despite the functional conservation within the family. In order to minimize the total number of probes while covering as many families as possible across a phylogenetically diverse set of organisms, we developed rigorous algorithms to choose optimal sets of conserved probes that ensured detection of divergent sequences within a family.

We note that the optimal characteristics of probes for gene family detection differ greatly from those for applications such as gene expression, in which mismatch bases are not tolerated and ideal probes produce linear signals in response to target concentration. For gene family detection, probes are required to tolerate a certain number of mismatches, commensurate with the degree of polymorphism within a gene family, without cross-hybridizing to members of other families. Linearity of response is not a concern, since our goal is to measure presence rather than abundance; in fact, we prefer to have probes that saturate in response to small quantities of complementary DNA. For this purpose, it worked well to calculate predicted free energies of hybridization for candidate probes against their complements, along with free energies for homodimer formation and self-hybridization. By setting a minimum threshold for an empirically derived linear combination of these free energies, we were able to select probes that had the necessary degree of sensitivity and mismatch tolerance.

Because environmental samples may contain limited quantities of intact pathogen DNA, we expect that sample material will need to be amplified to generate the amount of target DNA required to produce a detectable signal on the array. Whole genome amplification has been used widely for bacterial genomes, producing high yields with low bias. In a study by Arriola *et al* for comparative genome hybridization microarray of cancer samples, they have shown that the amplification biases using ϕ29 polymerase is less than 0.5% when sufficient material is used (Arriola *et al*, 2007). Wu *et al* have reported that they were able to detect as little as 10 fg of microbial community DNA on their 50-mer functional gene array when combined with whole community genome amplification (Wu *et al*., 2006). This amplification technique provides many advantages over specific amplification when the organism and mechanism to be identified are unknown. The amplified DNA can be directly labeled with fluorescent dyes and hybridized to NimbleGen arrays.

In our own limit of detection study, we applied whole genome amplification to initial quantities of fragmented *S. aureus* DNA as low as 0.31 fg and hybridized the amplified DNA to our NimbleGen array. We found we were able to detect 100% of the virulence and antibiotic resistance probes expected to be present in *S. aureus* in a sample amplified from 3.1 fg of starting DNA. Since the *S. aureus* strain used has a genome mass of 2.95 fg, this starting amount is equivalent to slightly more than one genome copy.

We used two approaches to assess the array's tolerance for mismatches between probe and target sequence. The first was to design more than 24,000 probes with mismatch nucleotides at known positions, and compare their performance to perfect match probes targeting the same bacterial sequences. We found that long oligomer probes with 50 to 66 nucleotides yielded greater than 90% detection rates, even with up to three mismatches from the target sequence. Probes with larger numbers of mismatches still gave high detection rates, provided there was a region of perfect match sequence with length at least 29 nt. A previous study by Kane *et al* has shown that a 50mer probe, which had a 15-base, 20-base or 35-base stretch matching with nontargets, had approximately 1%, 4% or 50% of the target signal intensity (Kane *et al*., 2000). He *et al* has reported that for a 70mer probe, when the stretch of PM probe sequence length is 20, 35 or 50 nt, the signal intensity reached 10%, 32% or 55% of the PM probes signal intensity (He *et al*, 2005). Our probes appear to have much better tolerance to stretches of mismatches. This could be due to our rigorous probe design algorithm.

Higher mismatch tolerance was seen when the mismatches were placed nearer the 3' end of the probe, which is anchored to the array surface. Conversely, the region of maximum sensitivity to mismatches is about 1/3 of the distance from the 5' end to the 3' end. This position dependence has been observed in other studies, but is not accounted for in current algorithms for prediction of hybridization free energies. Our probe design software for future generations of virulence gene detection arrays will factor in this dependence, placing more highly conserved regions of gene families in the areas of maximum impact along the length of the probe.

Our other approach to assess mismatch tolerance was to hybridize the array to genomic DNA from organisms with varying degrees of relatedness to the four target strains used for probe design. We found that a different *E. coli* strain from the two used for probe design still gave good results, with over 98% of gene families having detected probes. A *Salmonella* strain, which belongs to the same taxonomic family (*Enterobacteriae*) as *E. coli*, also had a large fraction of expected gene families (64%) with detection signals. Interestingly, the array performed less well with a member of the *Staphylococcus* genus, with only 47% of expected gene families being detected. These results may simply indicate that taxonomic categories are only a rough measure of phylogenetic relatedness. These preliminary results are encouraging, at any rate; they suggest that a future array with probes for each gene family sampled intelligently from the whole range of bacterial taxonomic families stands a reasonable chance of being able to detect orthologs from species that are not currently sequenced.

One of the limitations of functional gene arrays is that we cannot detect SNP-based or small indel-based mutations that affect virulence or resistance, because our array is designed to not be sensitive to small mutations, which is the cost of making the array broad enough to detect patterns of gene presence. Thus, it would be good to pair the virulence array with a tiling array for genes whose slight molecular variations are understood as to their affects on virulence or drug resistance, so the tiling array could be used as a secondary analysis if the primary virulence array indicated that a gene known to have important sequence variations was present.

In conclusion, the NimbleGen virulence gene array we developed shows great promise for detection of a broad range of virulence and antibiotic resistance genes. In addition to providing strain-level identification of known organisms, this technology will be valuable for functional characterization of unknown biothreat organisms. As a concrete example, we will use the array to identify or provide nearest-neighbor matches to organisms present in environmental samples collected by the BioWatch program (http://www.fas.org/sgp/crs/terror/RL32152.html), and to assess the pathogenic capabilities of unidentified organisms. Thus, our array can provide orthogonal confirmation for signature-based detection methods such as PCR. This array can also be used to differentiate virulent and avirulent strains by including antivirulence genes on the array (Maurelli and Prunier, 2007). Finally, the tools we have developed to design and analyze these arrays can be applied to create other kinds of functional gene arrays that will be valuable for the discovery of new pathogens, monitoring the metabolic capabilities of environmental microbes, and performing functional forensic analysis.

METHODS

Virulence gene sequence selection

Gene target sequences were selected from the genomes of the four bacterial strains shown in Table 1. These strains were selected because they were commercially available, have genome sequences published in GenBank (Benson *et al*. 2000), and required no more than a biosafety level 2 laboratory for sample processing. In addition, each has representatives of a wide variety of virulence-related gene families. Because our study focused on designing robust, sensitive probes for gene families, our target sequence set includes virulence gene orthologs found in strains such as *E. coli* K12 that (for poorly understood reasons) are avirulent to humans.

We selected gene sequences from the four genomes by searching with profile hidden Markov models (HMMs) for a set of 712 virulence-associated gene families gathered by Swan *et al*. [Swan 2005] from the literature and public databases. The HMMs found matches for 299 of the families in the four genomes, with some gene families represented by multiple paralogs in a given genome, and others by none, resulting in a total count of 1,245 gene sequences. The search was performed using the “estwisedb” algorithm of the Wise 2.0 software (Birney *et al*. 2004) on the Thunder supercomputer at LLNL (http://www.llnl.gov/pao/news/news_releases/2007/NR-07-04-05.html). For all the predicted gene sequences, existing gene annotations were downloaded from GenBank and correlated with the coordinates of the HMM matches.

HMM hits for the same gene family in multiple strains of the same organism had very similar if not identical sequences for all the cases examined.

Probe design for virulence gene target sequences

After selecting and extracting target gene sequences, we designed probes as diagrammed in **Figure 6**. We selected probes for a given gene family from the most conserved regions of sequences within that family, while ensuring that each target sequence had a minimum number of probes that were complementary to it. Using the most conserved regions enabled coverage of more sequences with fewer probes, and thus detection of more potential families on a single array, than simply tiling probes across each target sequence. We included additional probes for divergent sequences not captured by the conserved probes so that all known orthologs within the 4 genomes could be detected.

Probes were developed for one gene family at a time for target gene sequences from all four organisms. Candidate probes were generated using MIT's Primer3 software (Rozen and Skaletsky, 2000). Initial candidate selection was based on the rough predictions of the melting temperature T_m derived by Primer3 from the percent GC content and salt concentration. We next used a modified version of Unafold (Markham and Zuker, 2005) to make more accurate predictions of T_m and the minimum free energies of probe - target hybridization ($\Delta G_{complement}$), probe – probe hybridization ($\Delta G_{homodimer}$), and probe – self hybridization ($\Delta G_{hairpin}$). $\Delta G_{complement}$ is the predicted Gibb's free energy of hybridization of a probe with its reverse complement. While Unafold is a highly accurate program for ΔG and T_m prediction, it was too slow given our need to calculate T_m's and ΔG's for millions of candidate oligo probes. We created an accelerated version of Unafold that ran more than ten times faster by using more efficient data structures and caching thermodynamic parameter tables in memory rather than reloading them for each probe. The predicted ΔG's were then used to compute an aggregate "$\Delta G_{adjusted}$" (described below) for each candidate probe.

After this initial screening step, we removed duplicate probe sequences, and used a custom Perl program to cluster probe sequences into a minimal set of equivalence groups. Equivalence groups were defined so that all probes in a group were complementary to a known set of target sequences, with each target sequence represented by at least one equivalence group. When necessary, additional candidate probes were generated using relaxed selection criteria to ensure full coverage of the targets in a gene family.

The resulting set of candidates was then downselected, using a custom Python program, to produce a maximum of 30 probes per equivalence group, each capable of detecting multiple target sequences in a given family. The downselection program used an iterative ranking algorithm that favored probes having lower (more negative) $\Delta G_{adjusted}$, high conservation in orthologous genes of distantly related organisms, and maximum dispersal across the target gene sequence.

In addition, we included 2,000 positive control probes from the four genomes. These were designed to be distributed widely across each genome, and to range in length between 50-66 nt, in GC content between 40-60%, and in T_m between 71-91 °C.

Probe design parameter optimization

Detection of pathogens or gene families represented in a genomic DNA sample requires different probe design criteria than those used for gene expression, ChIP-chip or resequencing purposes. Each target type requires an appropriate balance between probe sensitivity and specificity. Ideally, all probes should be sensitive enough to detect DNA at single-copy concentrations. However, probes intended to distinguish organisms at the species or strain level must be designed to avoid cross-hybridization; while probes used to detect the presence of gene families should allow some degree of mismatch between the probe and target sequence, congruent with the range of sequence variation among orthologs within the family.

In order to determine the effect of design parameters on probe sensitivity and specificity, we constructed a prototype array containing several hundred probes for each member of ten gene families having representatives in all four target species. For each target gene, probes were selected spanning a wide range of design parameters. Probe lengths ranged from 30 to 66 nt, GC content from 40% to 60%, predicted T_m from 66°C to 91°C, $\Delta G_{complement}$ from –30 to –92 kcal/mol, $\Delta G_{homodimer}$ from –1.5 to -12 kcal/mol, and $\Delta G_{hairpin}$ from +1.8 to -6 kcal/mol. Prototype arrays were synthesized and hybridized to 4 μg of pure genomic DNA from one of the four species, as described below.

Figure 7 shows log signal intensities for probes targeting a typical gene family, in which DNA complementary to one set of probes (those for *E. coli* CFT073) was present in the hybridization mix, while DNA for another set of probes (for *E. faecalis)* was absent; thus the signal seen for *E. faecalis* probes is entirely due to nonspecific hybridization and other sources of background noise. We found that probes with lengths above 50 nt gave significantly better signals, with better separation from background, than lengths in the 30 to 45 nt range. The predicted melting temperature and $\Delta G_{complement}$ are strongly correlated with probe length, but not entirely determined by it. We performed linear regression fits to the log intensity against each of the probe design parameters, and multiple regressions against several combinations of parameters. Of the individual probe parameters we examined, the best predictor of intensity (i.e., the one with the smallest residual variance) was $\Delta G_{complement}$; the best multivariate predictor was a combination of $\Delta G_{complement}$, $\Delta G_{homodimer}$, and $\Delta G_{hairpin}$.

We observed that the relationship of log intensity to thermodynamic parameters such as $\Delta G_{complement}$ is nonlinear, and shows evidence of chemical saturation for the most sensitive probes. In order to incorporate saturation into our probe response model and find the combination of thermodynamic parameters that was the best predictor of probe sensitivity, we fit our data to a Langmuir isotherm curve (Burden *et al*, 2004) (**Figure 8**), parameterized by the $\Delta G_{complement}$, $\Delta G_{homodimer}$, and $\Delta G_{hairpin}$ as follows:

$$Intensity = a_0 + a_1/(1 + exp(a_2 + a_3\ \Delta G_{complement} + a_4\ \Delta G_{hairpin} + a_5\ \Delta G_{homodimer}))$$

We performed a nonlinear least squares fit to data from eight microarrays, each hybridized to 1-5 μg of DNA from one of the four target species, to fit values for the parameters a_0 through a_5. We determined that a linear combination of the three free energies which we term "$\Delta G_{adjusted}$" was the best predictor of hybridization intensity for probes complementary to the target DNA. The $\Delta G_{adjusted}$ is defined as:

$$\Delta G_{adjusted} = \Delta G_{complement} - 1.45\ \Delta G_{hairpin} - 0.33\ \Delta G_{homodimer}$$

In subsequent array designs, we screened candidate probes to include only those with predicted $T_m \geq 80$ °C, $\Delta G_{homodimer} > -12$ kcal/mol, $\Delta G_{hairpin} > -6$ kcal/mol, and $\Delta G_{adjusted} \leq -55$ kcal/mol.

Mismatch probe permutation methods

We generated mismatch probes, derived from a selection of perfect match target probe sequences, in order to test the ability of probes designed against gene family members from one organism to detect orthologs with non-identical sequences from other organisms. Previous experiments suggested that hybridization to mismatch probes depends strongly not only on the number of mismatched bases, but also on their location and distribution across the length of the probe. Mismatch probes were generated using five different strategies, incorporating single, adjacent, random, interval, and shifting perfect match region mismatches. Single and adjacent mismatch probes were generated by sliding a window of size *k* (with *k* taking values 1, 2, 3, 6, 10, 15 and 20) across the perfect match sequence, and creating *k* mismatches at the location of the window. We generated random mismatch probes by selecting *k* random positions in the perfect match probe, with *k* = 1, 2, 3, 6, 10, 15 or 20, and creating single mismatches at each position. In interval mismatch probes, mismatches were placed at regular intervals of size *k*, starting with a mismatch at the first base at the 5' end of the probe. Mismatch probes with shifting perfect match regions were created for region lengths *n* between 15 and 30, and offset values *s* ranging from 2 to 29. For each combination of length and offset, a probe was generated by preserving a perfect match region of size *n*, starting at base position *s*, and creating a mismatch at every third base on either side of the perfect match region.

Random control probe design

We generated 3,000 negative control probes, consisting of random sequences designed to have the same distribution of length and GC content as the target probes. BLAST searches of the random probes against the GenBank nt database showed that none had any perfect match alignments of length greater than 21 nt to any known sequence, so that none would be expected to hybridize to the organisms we tested on the arrays. These were used to determine background noise levels due to nonspecific hybridization of target DNA and to fluorescence of the chip substrate.

Microarray Synthesis

DNA microarrays were prepared on glass microscope slides according to a photolabile deprotection strategy that has been previously described [Singh-Gasson *et al*, 1999]. Arrays were generated at the LLNL MicroArray Center (LMAC). Reagents and supplies for the microarray syntheses were purchased from NimbleGen. Between 3 and 5 replicate features were generated for each probe. Using a checker-board pattern leaving every other spot vacant, 387,604 features were produced per array. The final deprotection and QC of the arrays were carried out as described [Nuwaysir *et al*, 2002]. Each array contained approximately 3,000 24-mer CPK6 (*Arabidopsis* calmodulin protein kinase 6) fiducial spots. The slides were hybridized with complimentary CPK6-Cy3 (Integrated DNA Technologies) and scanned to assure the quality of each array before hybridizing to DNA targets.

Sample preparation and microarray hybridization

E. coli K12 MG1655, *E. coli* CFT073, *E. faecalis* V583 and *S. aureus* Mu50 were purchased from ATCC. The bacterial culture pellets were grown according to the instructions from ATCC and genomic DNA was extracted using the Epicentre DNA extraction kit according to the manufacturer's protocols. The DNA was quantified using a NanoDrop spectrophotometer (Wilmington, DE). DNA samples were sonicated to fragment the DNA to a size range of 500-2000 bp, and then labeled with Cy3 labeled random nonamer primers (TriLink Biotechnologies, San Diego, CA) at 37^{o}C for 3 hr. The labeled samples were precipitated in isopropanol and the pellet washed, dried, reconstituted and quantified. For each hybridization, 4 μg of labeled DNA was mixed with Cy3- and Cy5-labeled CPK6 DNA, NimbleGen hybridization components and hybridization buffer according to the manufacturer's protocols. The NimbleGen arrays were hybridized with labeled DNA on a MAUI hybridization station (BioMicro Systems, Salt Lake City, UT) at 42^{o}C for 16 hr. Arrays were washed with NimbleGen wash buffers I, II and III according to NimbleGen protocols and scanned using an Axon GenePix 4000B scanner at 5 μm resolution.

For limit of detection experiments, aerosol filters were kindly supplied by the BioWatch program and DNA was extracted using the MoBio UltraClean Soil DNA kit (MoBio Laboratories, Carlsbad, CA), as described in (Radosevich *et al*, 2002). DNA from one filter was used as a common background, to which varying quantities of fragmented *S. aureus* DNA were added. *S. aureus* DNA was quantified using a PicoGreen double stranded DNA quantitation kit (Invitrogen, Carlsbad, CA), serially diluted, and then spiked into 10 ng of aerosol sample DNA in quantities of 0.31 fg, 3.1 fg, 31 fg, 310 fg or 3.1 pg. We performed whole genome amplification of the combined samples (aerosol samples with spiked-in *S. aureus* DNA, plus one pure aerosol sample) at 30^{o}C for 16 hr using the RepliG whole genome amplification kit (Qiagen, Valencia, CA). The amplified material was inactivated at 65^{o}C for 3 min and then purified using the QiaQuick PCR purification kit (Qiagen) to remove the primers and dNTPs. The entire amplified product was labeled with Cy3 random primer using the Klenow fragment and then hybridized to the array as described above.

For experiments on detection of virulence gene orthologs in related organisms, *E.coli* O157:H7 strain EDL933, *Staphylococcus saprophyticus* subsp. *saprophyticus strain* ATCC 15305, *Salmonella enterica* subspecies *enterica* serovar *Paratyphi A* strain ATCC 9150, and *Streptococcus pyogenes* strain MGAS5005 were purchased from ATCC. Samples were prepared from these organisms as described above for the other pure bacterial cultures.

Statistical methods for data analysis

Data were analyzed using custom software based on the R programming environment and BioConductor packages. Each probe was randomly spotted in three to five replicates to control for positional effects on the array. Data from replicate probes were summarized by the median of the $\log_2$-transformed intensities. Each probe on an array was considered to have a positive signal if the median $\log_2$ intensity of its technical replicates was above a detection threshold calculated for that array. The detection threshold was determined by using random control probes to model background noise. For each array, the detection threshold was set to the median $\log_2$ intensity of the random control probes, plus 4 times the standard deviation of the $\log_2$ intensities.

ACKNOWLEDGEMENTS

The authors thank Kat Swan for her diligent work in mining virulence and antibiotic resistance genes and building HMM models. The authors also thank Kris Montgomery for isolating the genomic DNAs used in this study and Thomas Bunt for providing the BioWatch aerosol DNA extracts. This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

FIGURE LEGENDS

Figure 1: Hybridization of *E. coli* CFT073 and *E. faecalis* to NimbleGen chips. 4µg of Cy3-labeled *E. coli* CFT073 or *E. faecalis* were hybridized to NimbleGen chip and the log2 intensity vs probe complement ΔG. Random control, *E. coli* CFT073, *E. coli* K12, *E. faecalis* and *S. aureus* probes were shown in red, yellow, green, cyan and purple colors. (A) *E. coli* CFT073. (B) *E. faecalis*.

Figure 2: Detection of virulence genes from related organisms. Data was plotted from the signal intensity vs the ID of msets (virulence gene families). Figure 2A and 2B shows gene families that are supposed to be present or absent in *E. coli* O157:H7 strain EDL933. Figure Figure 2C and 2D are for *S. saprophyticus* strain ATCC 15305. Figures 2E and 2F are for *S.enterica* serovar *paratyphi*. Figures 2G and 2H are for *S. pyogenes.*

Figure 3: Effect of the position of mismatches on the hybridization of target and probe. The mean mismatch probe intensity vs perfect match probe intensity ratio, averaged over 60 PM probes and their corresponding MM probes, from 10 arrays was plotted vs the position of the mismatch

Figure 4: Effect of the length of perfect match sequence on the hybridization of target and probe. 2, 3, 6, 10, 15 and 20 mismatches were randomly created. Intensity ratios were

averaged over 10 arrays for the 60 sets of PM and MM probes, and are plotted here on a log scale against the length of the longest PM region.

Figure 5: Detection of virulence genes from *S. aureus* spiked in BioWatch aerosol filter samples. 0.31 fg, 3.1 fg, 31 fg, 310 fg and 3.1 pg of *S. aureus* were spiked into 10 ng of extracted BioWatch aerosol samples. The samples were amplified, labeled and hybridized. Log2 intensity of probes vs complement ∆G was plotted.

Figure 6: NimbleGen array probe design process. Candidate probes were generated using Primer3 and Unafold based on T_m, GC content, salt concentration, and minimum free energies of probes. Probes were filtered based on best free energy and duplicate probe sequences were removed. When necessary, additional candidate probes were generated using more relaxed parameters to ensure full coverage. The final set of probes was then downselected to produce a maximum of 30 probes per equivalence group, each capable of detecting multiple target sequences in a given family.

Figure 7: Log2 intensity vs probe length (A), predicted melting temperature *Tm* (B), and predicted complement ∆G (C) for selected probes in an array hybridized with *E. coli* genomic DNA. Probes specific for *E. coli* sequences are plotted in green; probes specific for *E. faecalis* are in red.

Figure 8: Langmuir isotherm fit to example data from one array.

Supplementary table S1: List of gene families present or absent from Figures 2A-H. The symbols in each organism column consist of two slash-separated characters: first a plus or minus according to whether the gene family has an ortholog or not in the given species; then a plus, zero or minus according to whether the ortholog was detected (signal above brightest random control), marginal (signal between median of controls + 4 SD and the maximum control signal), or undetected (signal below median + 4 SD).

Table 1: Bacterial Strains Used for Probe Selection

Table 2: Input parameters for Primer3 probe generation

REFERENCES

Gentry,T.J., Wickham,G.S., Schadt,C.W., He,Z. and Zhou,J. 2006. Microarray Applications in Microbial Ecology Research. *Microbial Ecology* **52**, 159–175.

Rhee,S.K., Liu,X., Wu,L., Chong,S.C., Wan,X. and Zhou,J. 2004. Detection of genes involved in biodegradation and biotransformation in microbial communities by using 50-mer oligonucleotide microarrays. *Appl Environ Microbiol* **70**: 4303–4317.

He,Z, Gentry,T.J., Schadt,C.W., Wu,L., Liebich,J., Chong,S.C., Huang,Z., Wu,W., Gu,B., Jardine,P., Criddle,C., and Zhou,J. 2007. Geochip: a comprehensive microarray for investigating biogeochemical, ecological and environmental processes. *The ISME Journal*, **1**, 67-77.

Palacios,G., Quan,P., Jabado,O.J., Conlan,S., Hirschberg,D.L., Liu,Y., Zhai,J., Renwick,N., Hui,J., Hegyi,H., Grolla,A., Strong,J.E., Towner,J.S., Geisbert,T.W., Jahrling,P.B., Büchen-Osmond,C., Ellerbrok,H., Sanchez-Seco,M.P., Lussier,Y., Formenty,P., Nichol,S.T., Feldmann,H., Briese,T. and Lipkin,W.I. 2007. Panmicrobial Oligonucleotide Array for Diagnosis of Infectious Diseases. *Emerging Infectious Disease*, **13**, 73-81.

Singh-Gasson,S., Green,R.D., Yue,Y., Nelson,C., Blattner,F., Sussman,M.R. and Cerrina,F. 1999. Maskless fabrication of light-directed oligonucleotide microarrays using a digital micromirror array *Nat. Biotech*., **17**, 974-978.

Benson,DA., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J., Rapp,B.A. and Wheeler,D.L. 2000. GenBank. *Nucleic Acids Res.*, **28**, 15-18.

Swan,K.A., Kuczmarski,T.A., Gardner,S.N., Slezak,T.R. 2005. Biopathogen virulence gene identification. Poster presented at the Annual IC Postdoctoral Research Fellowship Colloquium, Washington, DC, April 2005.

Birney,E., Clamp,M., Durbin,R. 2004. GeneWise and Genomewise. *Genome Res.,* **14***:* 988-995.

Rozen,S. and Skaletsky,H. 2000. Primer3 on the WWW for general users and for biologist programmers. In: *Bioinformatics Methods and Protocols: Methods in Molecular Biology* (ed. S. Krawetz and S. Misener, S), pp. 365-386. Humana Press, Totowa, NJ,.

Markham,N.R. and Zuker,M. 2005. DNAMelt web server for nucleic acid melting prediction. *Nucleic Acids Res*., **33**, W577-W581.

http://www.computerworld.com.au/index.php?id=1811327359&fp=16&fpid=0, 'Thunder'-- North America's fastest Linux supercomputer LinuxWorld staff (LinuxWorld) 14/05/2004 12:08:13

Burden,C.J., Pittelkow,Y.E. and Wilson,S.R. 2004. Statistical Analysis of Adsorption Models for Oligonucleotide Microarrays. In *Statistical Applications in Genetics and Molecular Biology*. Vol. 3 : Iss. 1, Article 35.
Available at: http://www.bepress.com/sagmb/vol3/iss1/art35

Nuwaysir,E.F., Huang,W., Albert,T.J., Singh,J., Nuwaysir,K., Pitas,A., Richmond,T., Gorski,T., Berg,J.P., Ballin,J., McCormick,M., Norton,J., Pollock,T., Sumwalt,T., Butcher,L., Poeter,D., Mola,M., Hall,C., Blattner,F., Sussman,M.R., Wallace,R.L., Cerrina,F. and Green,R.D., 2002. Gene Expression Analysis Using Oligonucleotide Arrays Produced by Maskless Photolithography. *Genome Res.*, **12**, 1749-1755.

Wilson,K.H., Wilson,W.J., Radosevich,J.L., DeSantis,T.Z., Viswanathan,V.S., Kuczmarski,T.A. and Andersen,G.L. 2002. High Density Microarray of Small-Subunit Ribosomal DNA Probes. *Applied and Environmental Microbiology*, **68,** 2535-2541.

Radosevich,J.L., Wilson,W.J., Shinn,J.H., DeSantis,T.Z. and Anderson,G.L. 2002. Development of a high-volume aerosol collection system for the identification of air-borne micro-organisms *Let. in Appl. Micro.*, **34**, 162-167.

He,Z., Wu,L., Li,X., Fields,M.W. and Zhou,J. 2005. Empirical establishment of oligonucleotide probe design criteria. *Appl. Environ. Microbiol,* **71**, 3753-3760.

Kane,M.D., Jatkoe,T.A., Stumpf,C.R., Lu,J., Thomas, J.D. and Madore,J.M. 2000. Assessment of the specificity and sensitivity of oligonucleotide (50mer) microarrays. *Nucleic Acid Res*. **28**, 4552-4557.

Arriola,E., Lambros,M.BK., Jones,C., Dexter,C., Mackay,A., Tan,D.SP., Tamber,N., Fenwick,K., Ashworth,A., Dowsett,M. and Reis-Filho,J.S. 2007. Evaluation of Phi29-based whole-genome amplification for microarray-based comparative genomic hybridization. *Lab. Investigation,* **87**, 75-83.

Wu,L., Liu,X., Schadt,C.W. and Zhou,J. 2006. Microarray-based analysis of sub-nanogram quantities of microbial community DNAs using Whole Community Genome Amplification (WCGA). *Appl. Environ. Microbiol.*, **72**, 4931-4941.

Maurelli, AT and Prunier, AL. 2007. Mutations, black holes, and antivirulence genes. *Microbe* **2**, 388-394.

Jackson,S.A., Mammel,M K., Patel I.R., Mays, T., Albert,T.J., LeClerc, J. E. Cebula, T.A. Interrogating genomic diversity of E. coli O157:H7 using DNA tiling arrays 2007. *Forensic Sci. International.* **168,** 183–199.

Figure 1

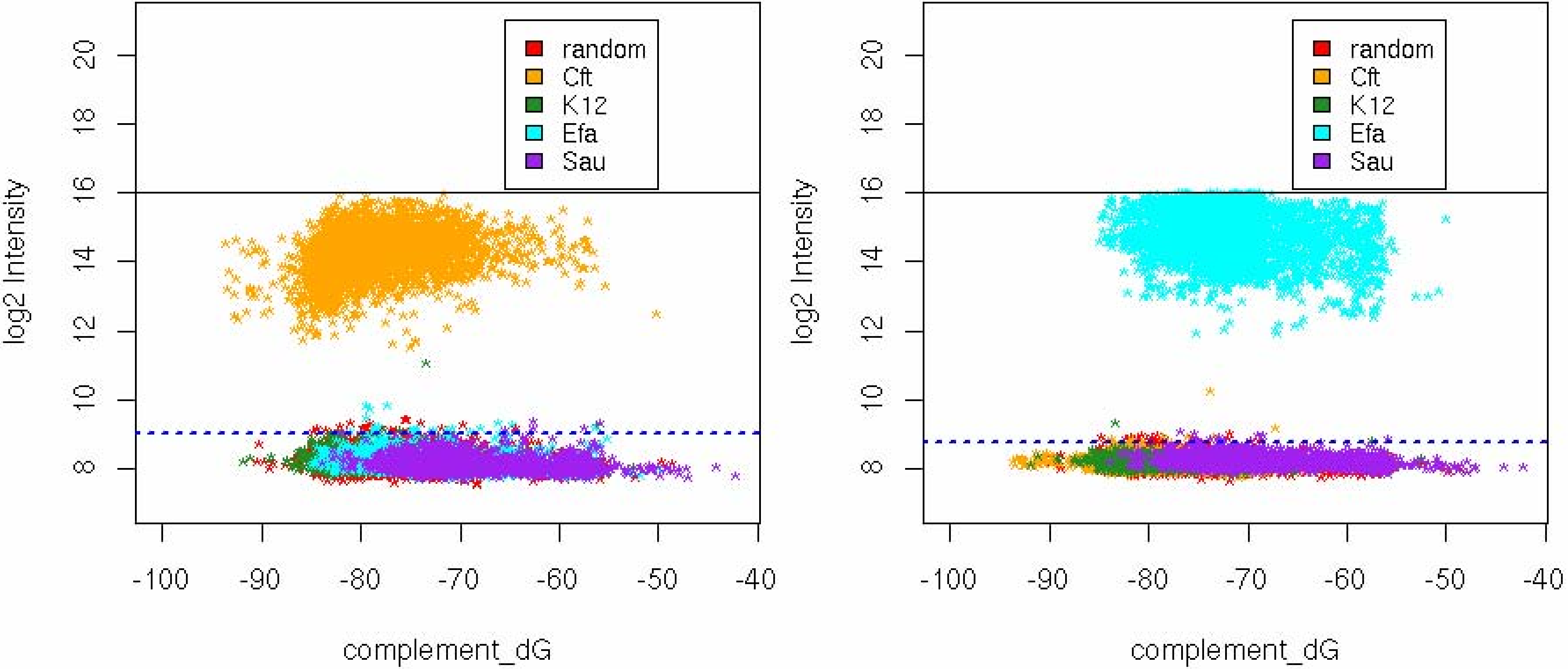

random
Cft
K12
Efa
Sau
log2 Intensity
8
10
12
14
16
18
20
-100
-90
-80
-70
-60
-50
-40
complement_dG
random
Cft
K12
Efa
Sau
log2 Intensity
8
10
12
14
16
18
20
-100
-90
-80
-70
-60
-50
-40
complement_dG

Figure 2 Legend

- 🟥 random
- 🟦 E_faecalis_V583
- ▢ E_coli_CFT073
- 🟩 E_coli_K12_MG1655
- 🟪 S_aureus_Mu50

Figure 2A

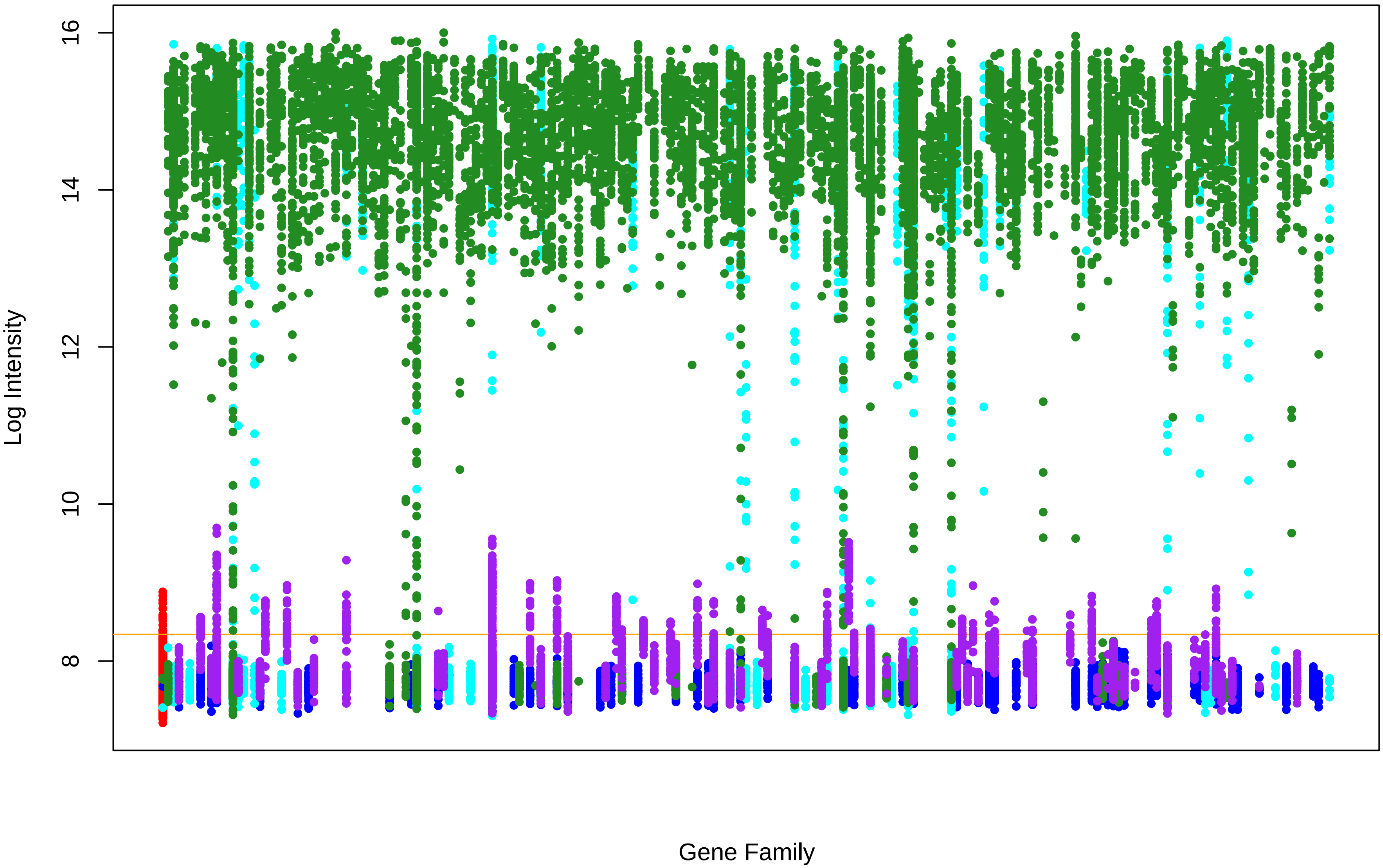


Figure 2B

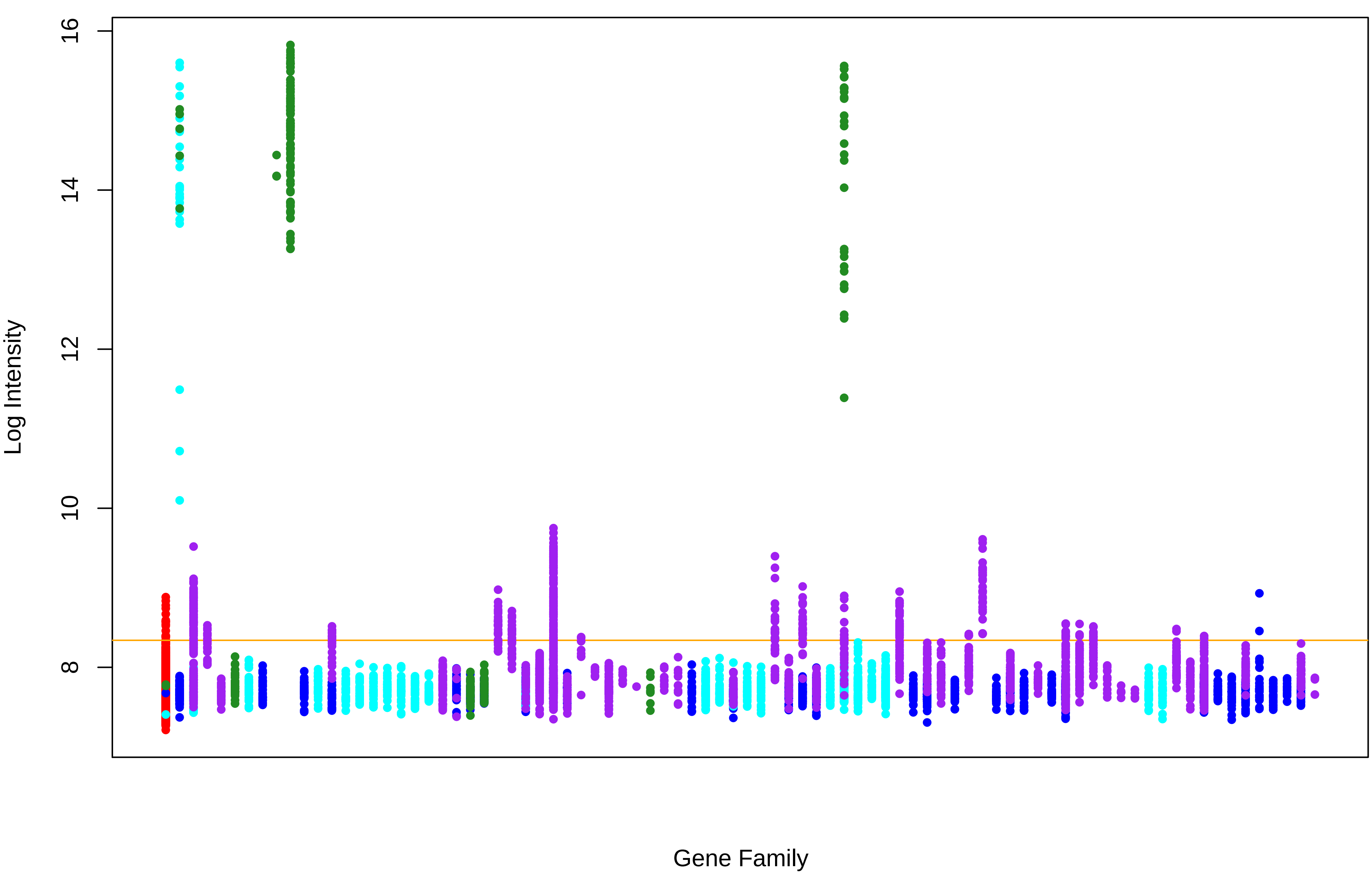


Figure 2C

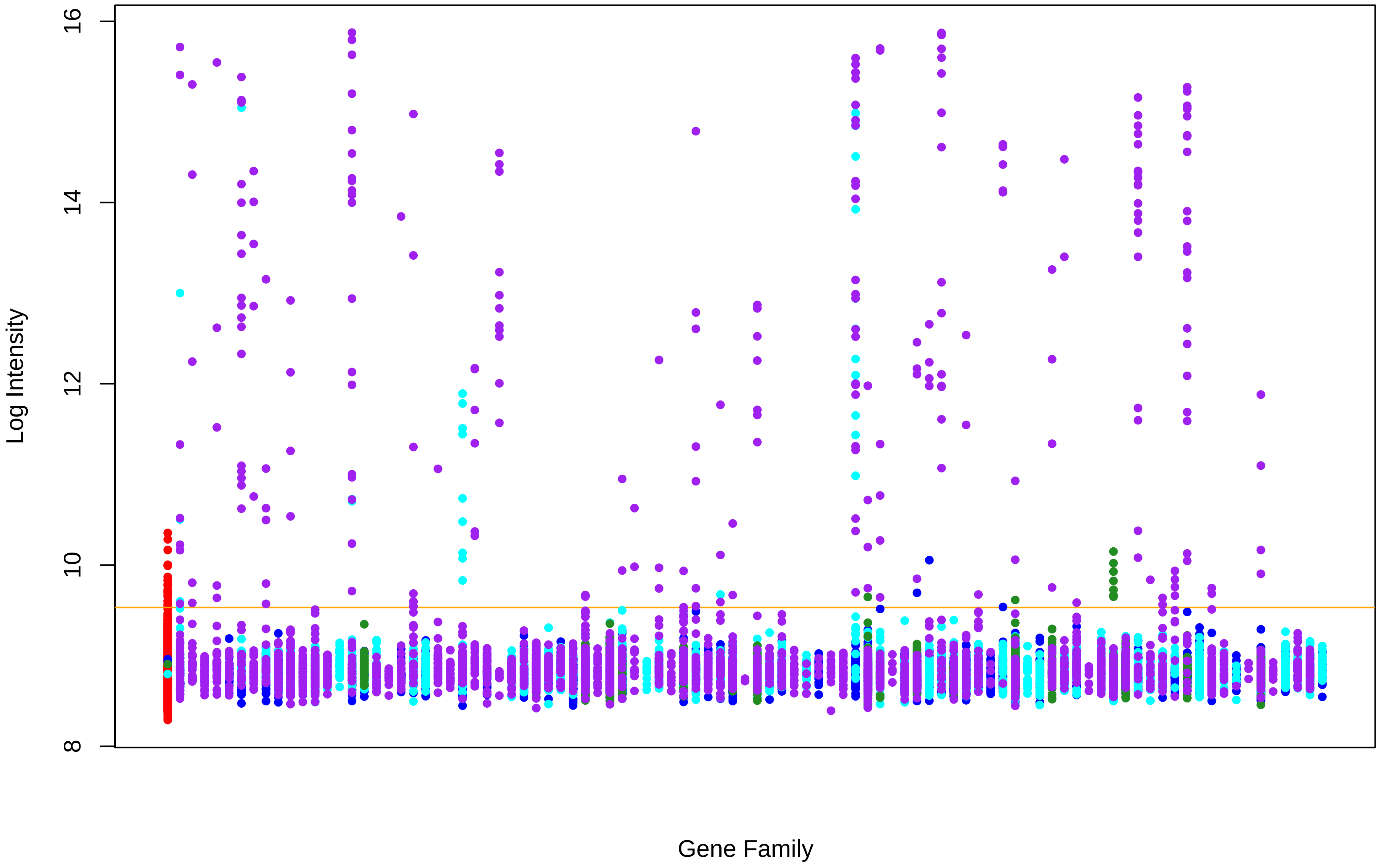


Figure 2D

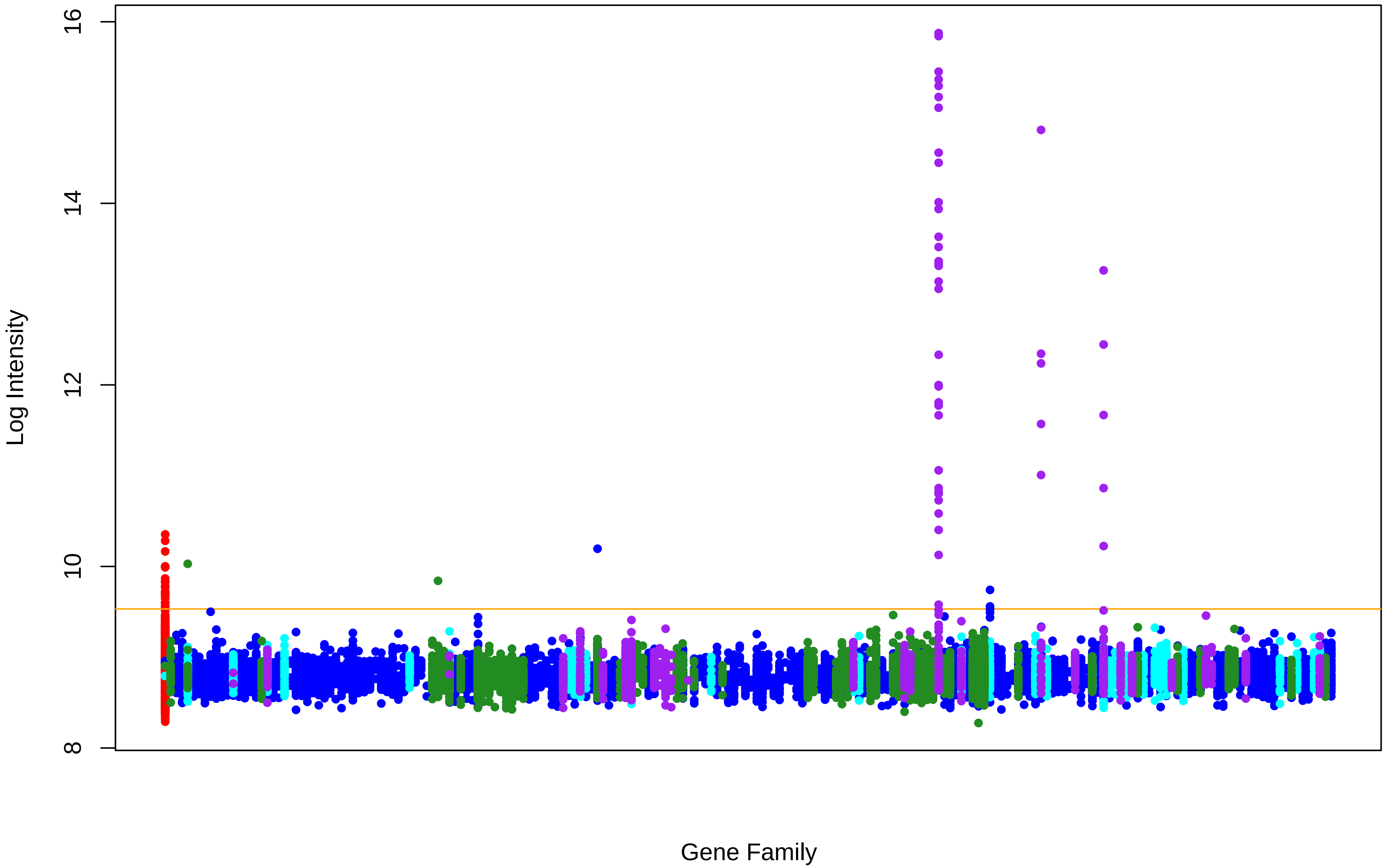


Figure 2E

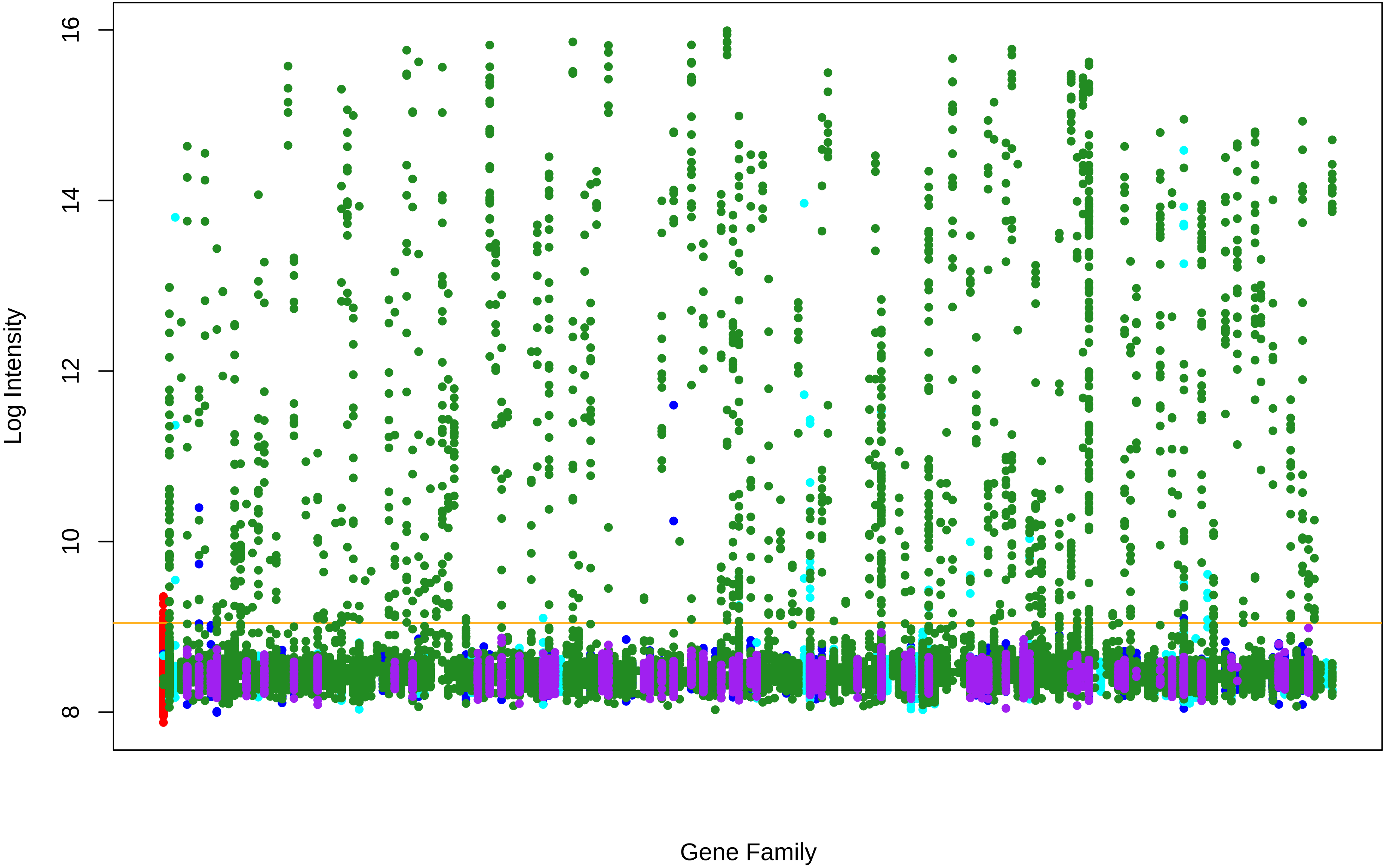


Figure 2F

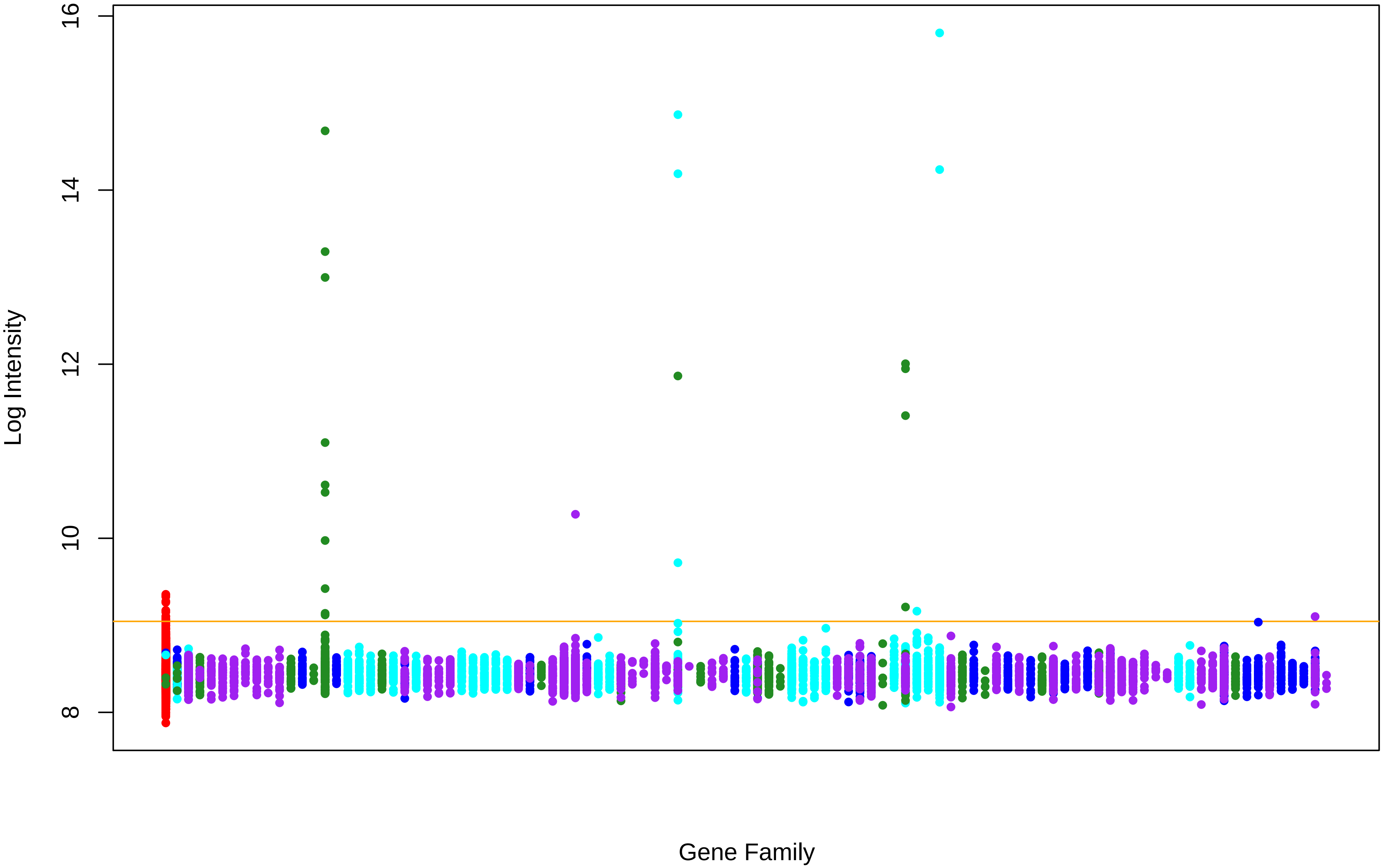


Figure 2G

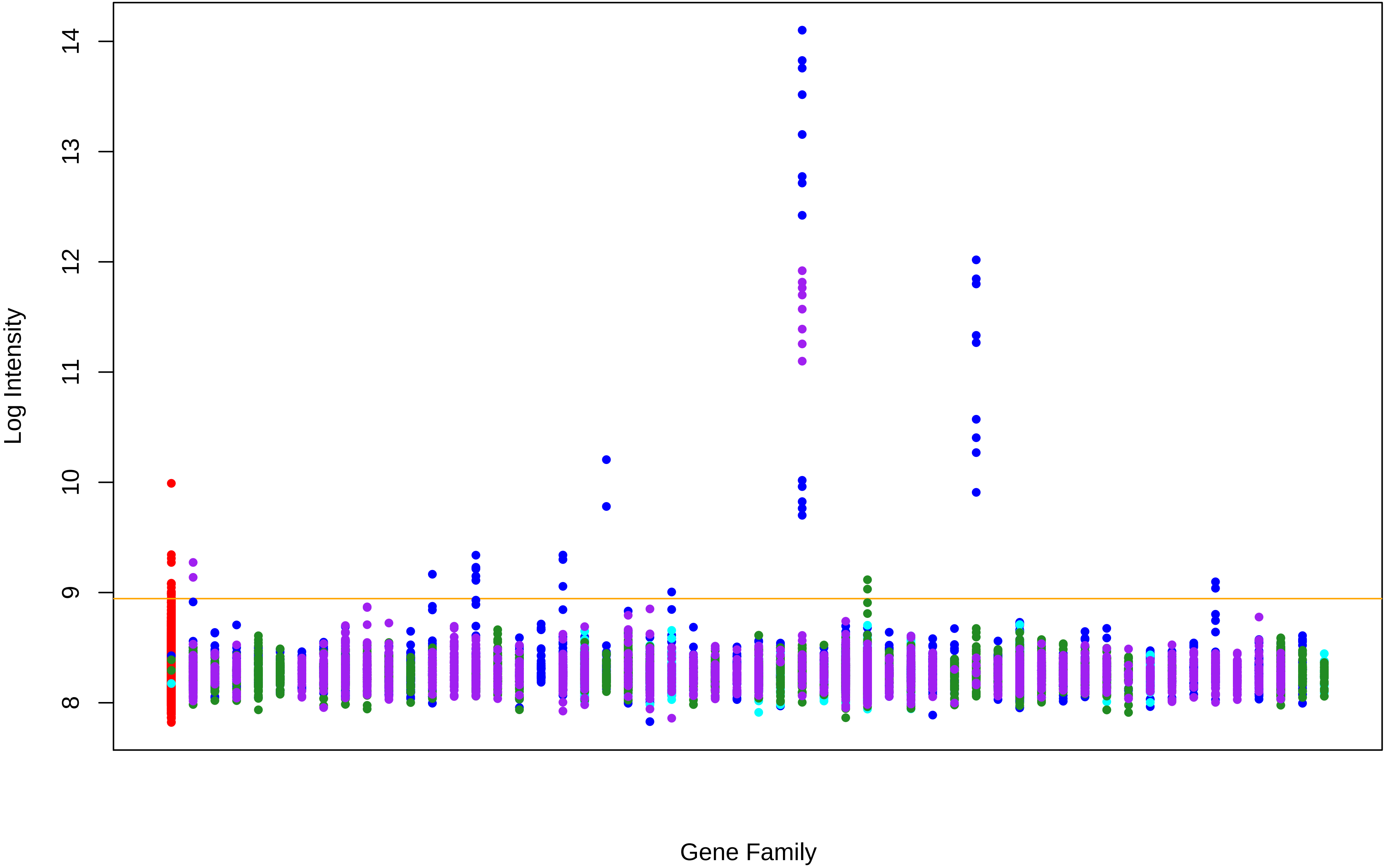


Figure 2H

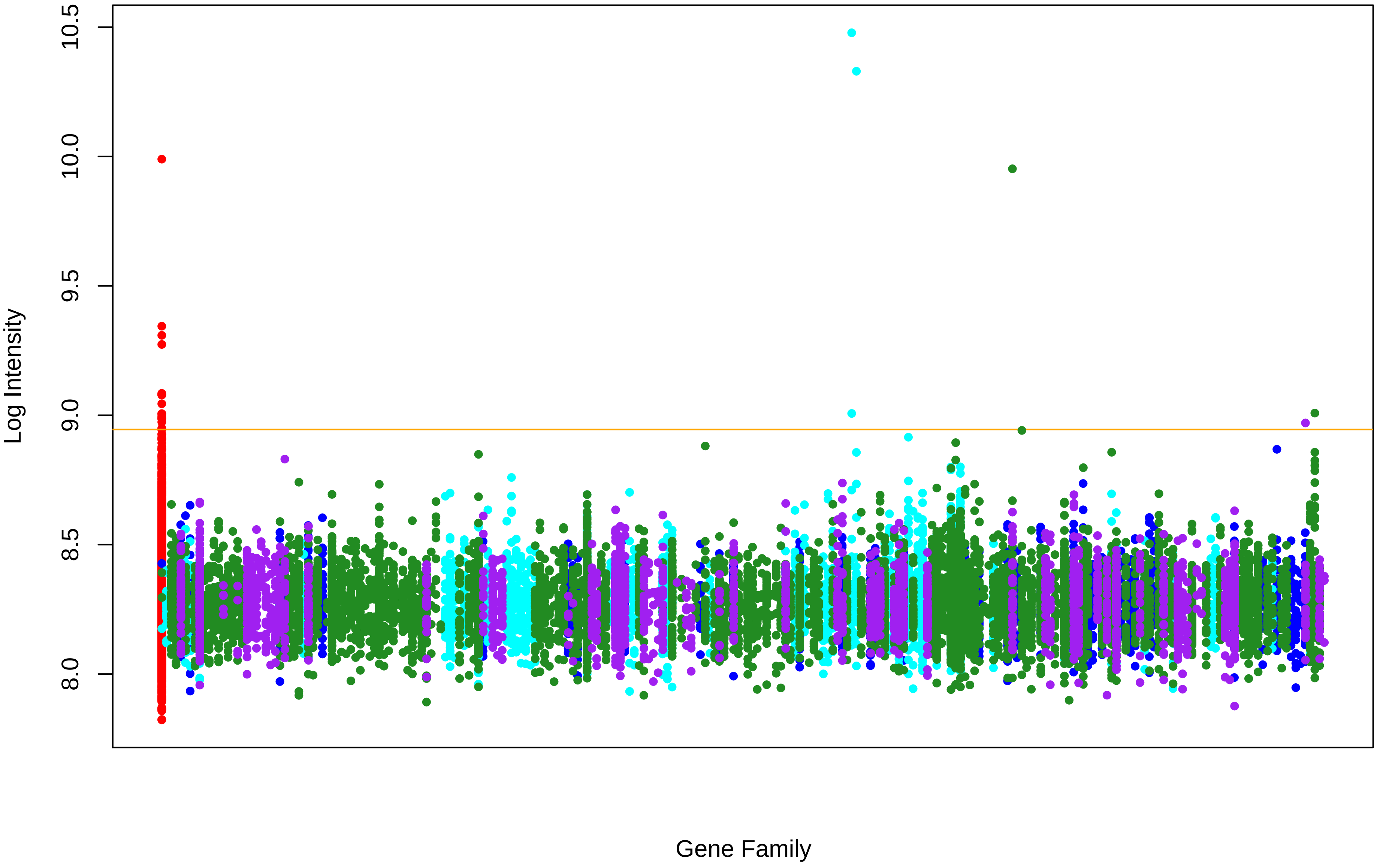


Figure 3

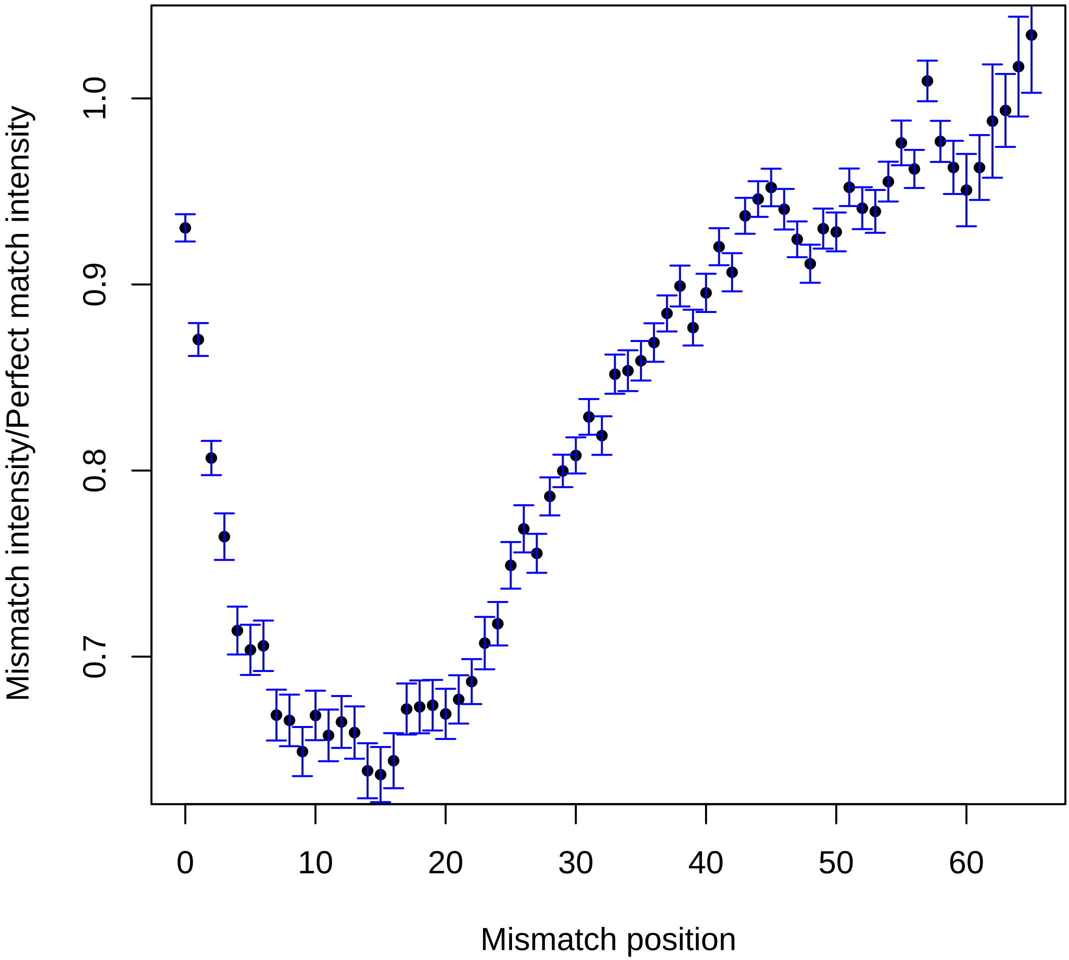
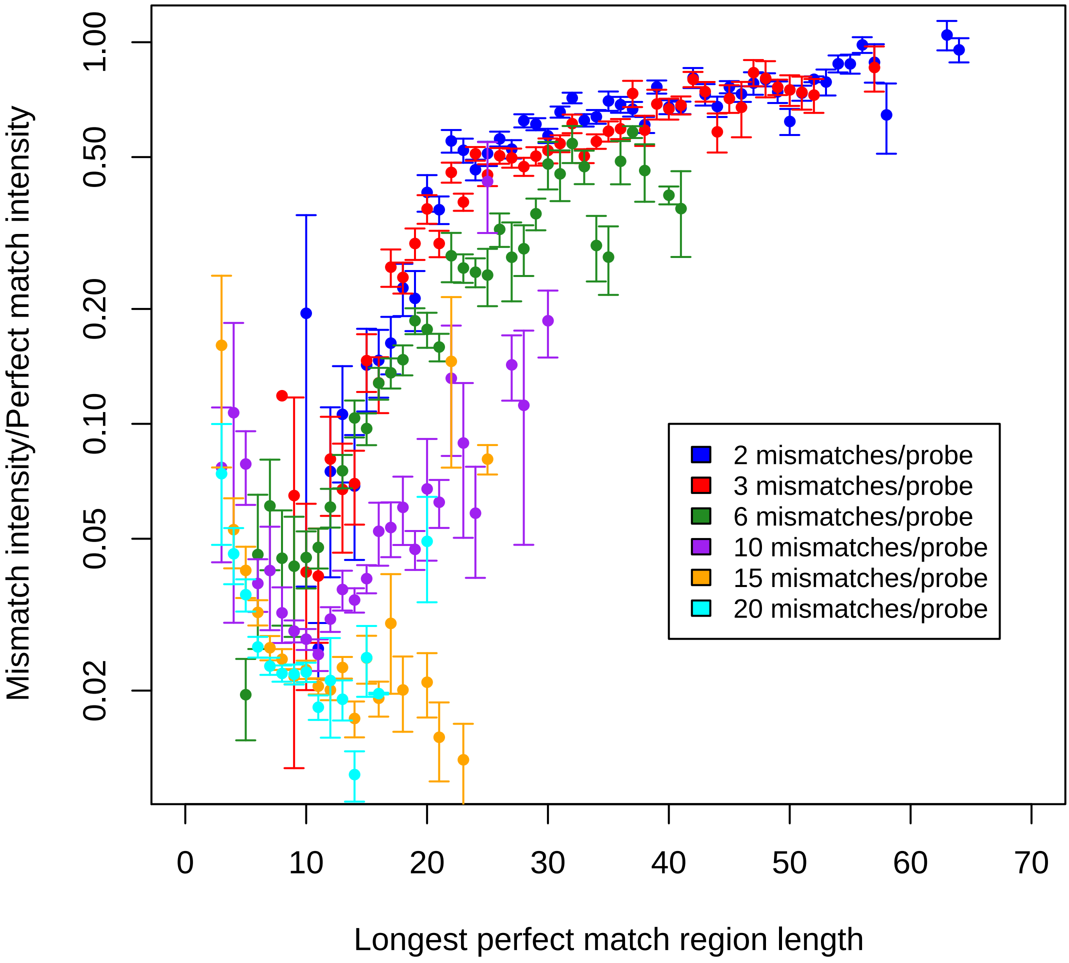
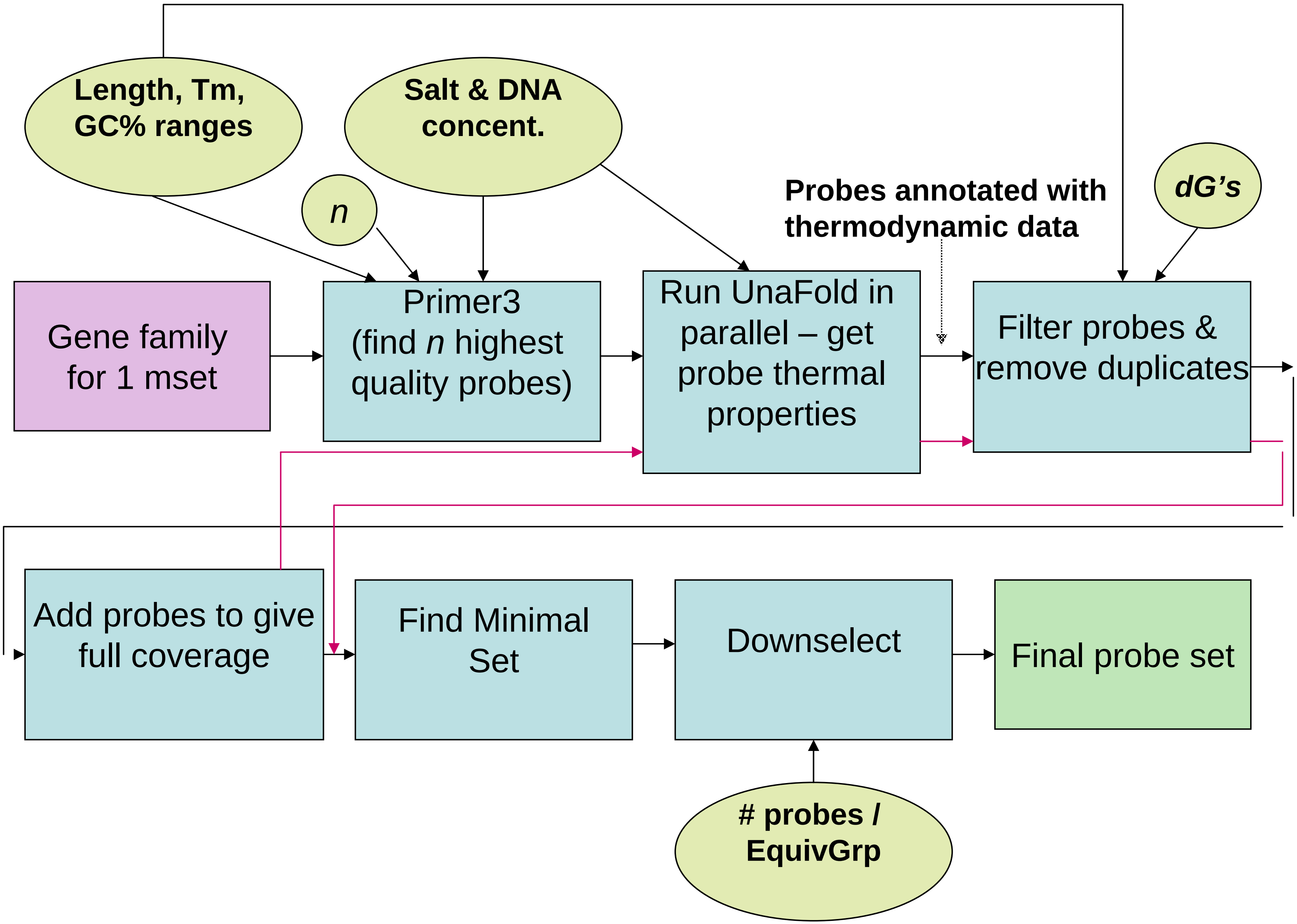

Mismatch intensity/Perfect match intensity
0.7
0.8
0.9
1.0
0
10
20
30
40
50
60
Mismatch position

Figure 4



Mismatch intensity/Perfect match intensity
1.00
0.50
0.20
0.10
0.05
0.02
0
10
20
30
40
50
60
70
Longest perfect match region length
🟦 2 mismatches/probe
🟥 3 mismatches/probe
🟩 6 mismatches/probe
🟪 10 mismatches/probe
🟧 15 mismatches/probe
🟦 20 mismatches/probe

Figure 6


Length, Tm, GC% ranges
Salt & DNA concent.
n
Probes annotated with thermodynamic data
dG's
Gene family for 1 mset
Primer3 (find n highest quality probes)
Run UnaFold in parallel – get probe thermal properties
Filter probes & remove duplicates
Add probes to give full coverage
Find Minimal Set
Downselect
Final probe set
# probes / EquivGrp

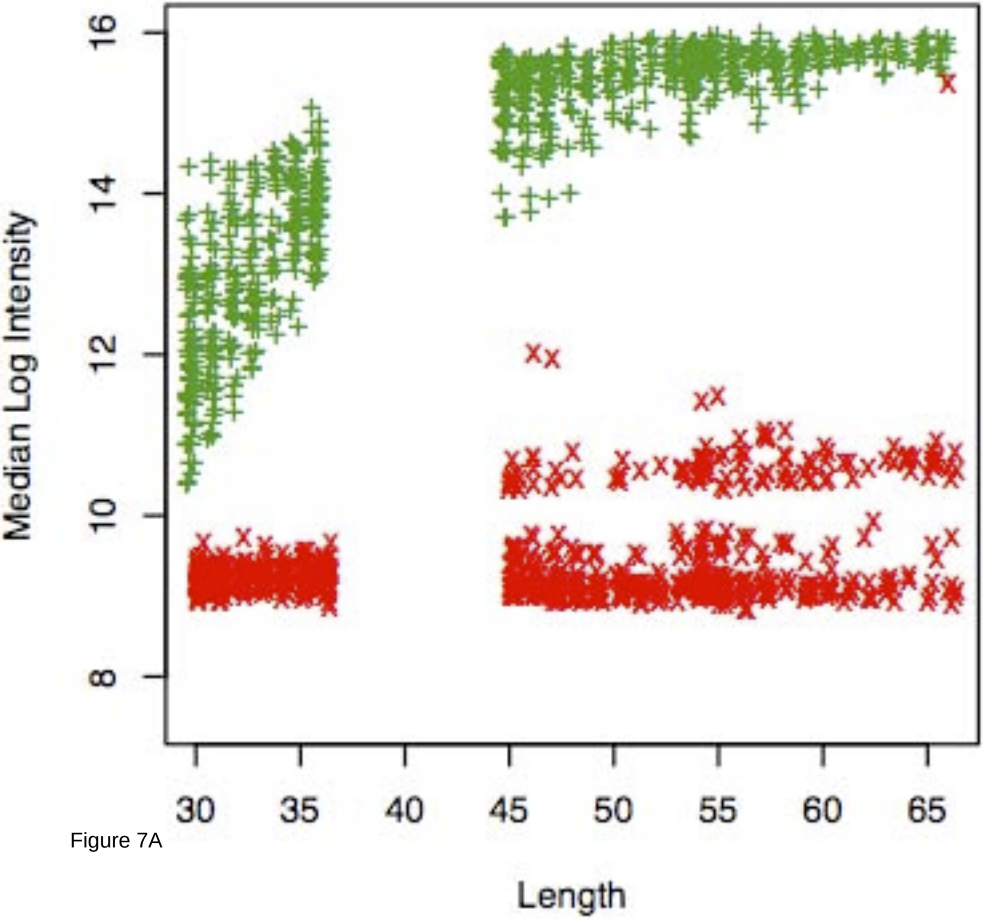


Figure 7A

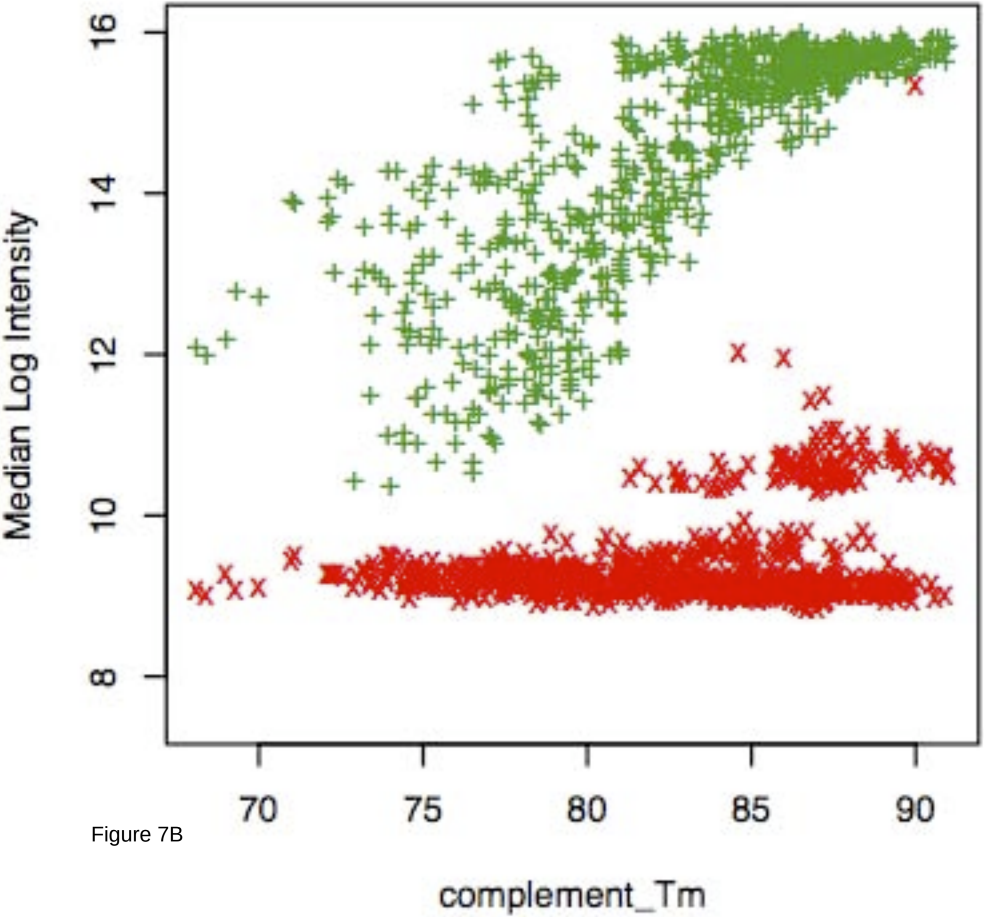


Figure 7B

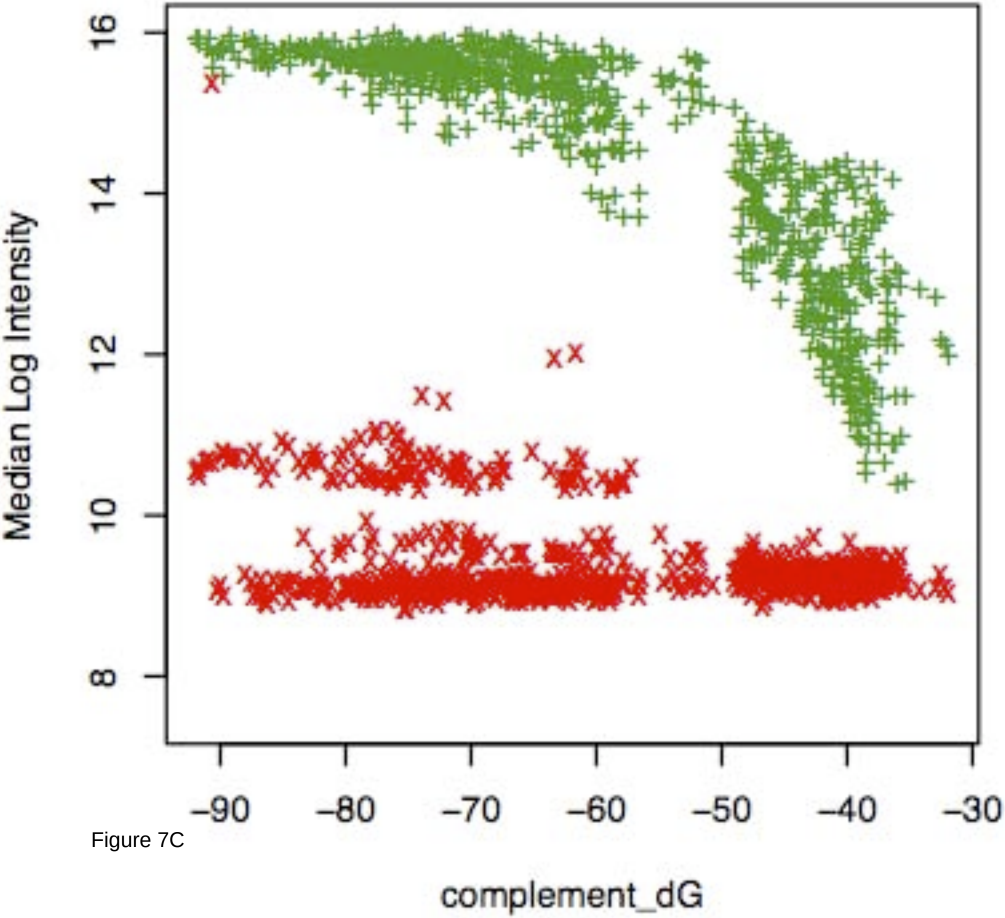

16
14
12
10
8
Median Log Intensity
−90
−80
−70
−60
−50
−40
−30
complement_dG

Figure 7C

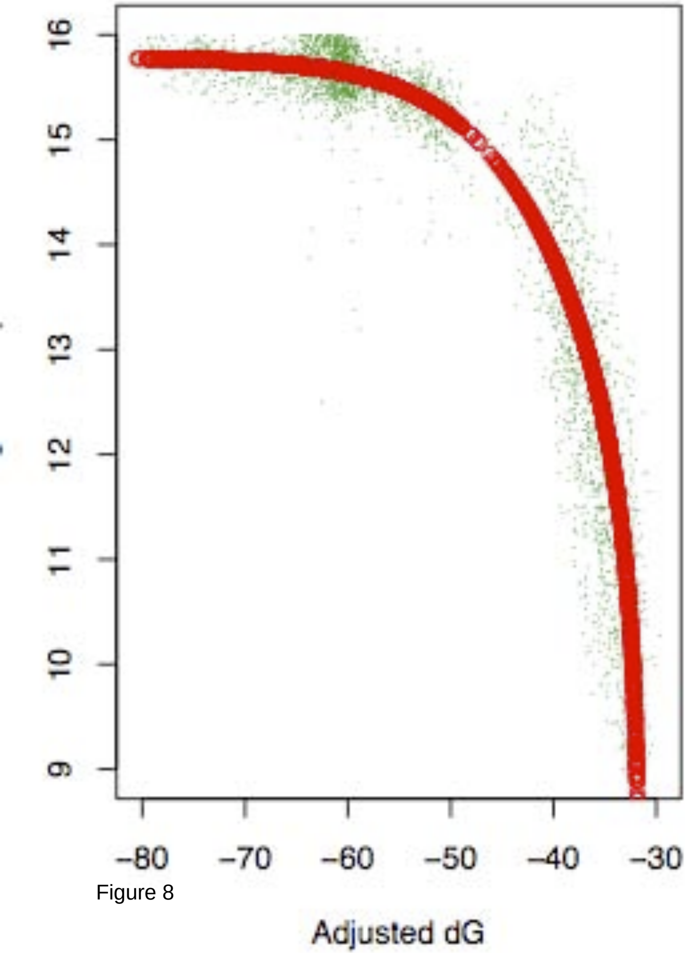


Figure 8

Supplementary Table 1

Family ID	Description	Category	Ortholog presence expected / detected by strain			
			E. coli O157 H7 EDL933	S. saprophyticus ATCC 15305	S. paratyphi ATCC 9150	S. pyogenes MGAS5005
2499	hemagglutinin (Hep_Hag) repeat protein	adhesion	+/+	-/-	+/+	-/+
2501	hemagglutinin (Hep_Hag) repeat protein, autotransporter	adhesion	+/+	-/-	+/+	-/+
8377	aad-6 Aminoglycoside 6-adenylyltransferase (EC 2.7.7.-)	antibiotic resistance	-/-	+/-	-/-	-/-
8319	aad-9 Streptomycin 3-adenylyltransferase (EC 2.7.7.47)	antibiotic resistance	-/0	-/-	+/-	-/-
8349	aminoglycoside phosphotransferase	antibiotic resistance	-/-	-/-	-/-	-/-
8339	ampC Beta-lactamase precursor (EC 3.5.2.6)	antibiotic resistance	+/+	-/-	-/-	-/-
8411	arnA Bifunctional polymyxin resistance arnA protein	antibiotic resistance	+/+	-/-	+/-	-/-
8415	bceA part of the ABC transporter complex bceAB (TC 3.A.1.123.5), bacitracin export	antibiotic resistance	-/0	+/-	-/-	+/0
8417	bceB Bacitracin export permease protein, part of the ABC transporter complex bceAB (TC 3.A.1.123.5)	antibiotic resistance	-/0	+/-	-/-	-/-
8335	bcr bicyclomycin resistance protein	antibiotic resistance	+/+	+/+	+/+	-/-
8361	blaR1, mecR1 antirepressor	antibiotic resistance	-/-	-/-	-/-	-/-
8439	ble bleomycin resistance protein	antibiotic resistance	-/0	-/-	-/-	-/-
8441	bll1, blo(2,3,5,7,B,F,K), blp2 beta-lactamase precursor (EC 3.5.2.6)	antibiotic resistance	-/-	-/-	-/-	-/-
2564	BPSS2119, putative metallo-beta-lactamase	antibiotic resistance	-/+	-/+	-/-	-/-
8321	cmlA chloramphenicol resistance	antibiotic resistance	+/+	-/-	+/+	-/-
8258	cmlR chloramphenicol resistance protein	antibiotic resistance	+/+	+/-	+/-	-/-
8308	dfrA Dihydrofolate reductase (EC 1.5.1.3)	antibiotic resistance	+/+	+/-	+/+	+/-
8459	fos(A,B) glutathione transferase fosA (EC 2.5.1.18), metallothiol transferase fosB (EC 2.5.1.-)	antibiotic resistance	-/-	+/-	-/-	-/-
8461	fosX fosfomycin resistance protein	antibiotic resistance	-/-	+/-	-/-	-/-
8463	fsr fosmidomycin resistance protein	antibiotic resistance	+/+	-/-	+/+	-/-
8465	hcp(A,D) Beta-lactamase hcp precursor (EC 3.5.2.6)	antibiotic resistance	-/-	-/-	-/-	-/-
8467	hcp(B,C,E) Putative beta-lactamase hcp precursor (EC 3.5.2.6)	antibiotic resistance	-/-	-/-	-/-	-/-
8471	kanU kanamycin nucleotidyltransferase (EC 2.7.7.-)	antibiotic resistance	-/0	-/-	-/-	-/-
8345	ksgA, erm(1,A,B,C) Dimethyladenosine transferase (EC 2.1.1.-) , rRNA adenine N-6-methyltransferase (	antibiotic resistance	-/-	+/-	-/-	+/0
8481	lmrB Lincomycin resistance protein lmrB	antibiotic resistance	-/0	+/+	-/-	-/-
8483	lpxD UDP-3-O-[3-hydroxymyristoyl] glucosamine N-acyltransferase (EC 2.3.1.-)	antibiotic resistance	+/+	-/-	+/+	-/-
8246	marB	antibiotic resistance	+/+	-/-	-/-	-/-
8284	marB	antibiotic resistance	+/+	-/-	+/-	-/-
8541	marC multiple antibiotic resistance protein	antibiotic resistance	+/+	-/-	+/+	-/0
8288	mdfA, mdtA	antibiotic resistance	+/+	-/-	+/-	-/-
8485	mdtH confers resistance to norfloxacin and enoxacin.	antibiotic resistance	+/+	-/-	+/+	-/-
8487	mdtN tripartite efflux system composed of mdtN, mdtO and mdtP	antibiotic resistance	+/+	-/-	+/+	-/-
8489	mdtO tripartite efflux system composed of mdtN, mdtO and mdtP	antibiotic resistance	+/+	-/-	-/-	-/-
8491	mdt(P,Q) tripartite efflux system composed of mdtN, mdtO and mdtP	antibiotic resistance	+/+	-/-	+/+	-/-
8280	mreB actin-like ATPase involved in rod shape-determining	antibiotic resistance	+/+	-/-	+/+	-/-
8495	msrA erythromycin resistance ATP-binding protein	antibiotic resistance	-/-	-/-	-/-	-/-
8497	myrA mycinamicin-resistance protein	antibiotic resistance	+/+	-/-	+/0	-/-
8533	pbp beta-lactam-inducible penicillin-binding protein	antibiotic resistance	-/-	-/-	-/0	-/0
8503	pmrD polymyxin B resistance protein	antibiotic resistance	+/+	-/-	+/-	-/-
8266	rarD chloramphenicol-sensitive protein	antibiotic resistance	+/+	+/-	+/-	-/-
8331	tcr tetracycline resistance protein	antibiotic resistance	-/+	-/-	-/-	-/-
8513	tehA Tellurite resistance protein	antibiotic resistance	+/+	-/-	+/+	-/-
8515	tehB Tellurite resistance protein	antibiotic resistance	+/+	-/-	+/-	-/-
8511	tet347 tcr Tetracycline resistance determinant	antibiotic resistance	-/-	-/-	-/-	-/-
8317	tetA(A,B,C,D,E,G,H)	antibiotic resistance	-/-	-/-	-/-	-/-
8270	tet(M,O,P,Q,S,W)	antibiotic resistance	-/-	-/-	-/-	-/-

8311	tetV, protein has 10 TM domains, possible efflux pump, possible drug aniporter enabling efflux of te	antibiotic resistance	+/+	+/0	+/+	-/-
8519	tlrC tylosin resistance ATP-binding protein	antibiotic resistance	-/-	+/-	-/-	+/-
8337	tmrB Tunicamycin resistance protein	antibiotic resistance	-/-	-/-	-/-	-/-
8256	uppp Undecaprenyl-diphosphatase (EC 3.6.1.27)	antibiotic resistance	+/+	+/0	+/+	+/-
8262	uppp Undecaprenyl-diphosphatase (EC 3.6.1.27)	antibiotic resistance	+/+	+/-	+/+	+/-
8264	uppp Undecaprenyl-diphosphatase (EC 3.6.1.27)	antibiotic resistance	+/+	+/0	+/+	+/-
8282	uppp Undecaprenyl-diphosphatase (EC 3.6.1.27)	antibiotic resistance	+/+	+/0	+/+	+/-
8304	uppp Undecaprenyl-diphosphatase (EC 3.6.1.27)	antibiotic resistance	+/+	+/+	+/+	+/-
8371	uppp Undecaprenyl-diphosphatase (EC 3.6.1.27)	antibiotic resistance	+/+	-/-	+/+	-/-
8387	uppp Undecaprenyl-diphosphatase (EC 3.6.1.27)	antibiotic resistance	+/+	+/+	+/+	+/-
8521	van(A,B) vancomycin/teicoplanin A-type resistance protein vanA or vanB (EC 6.3.2.-)	antibiotic resistance	-/+	-/-	-/-	-/-
284	ecotin, protease inhibitor	anti-elastin	+/+	-/-	+/-	-/-
7292	NDVA, beta(1,2) glucan export ATP binding protein	beta glucan export	-/-	-/-	-/-	-/-
2263	hmsF, pgaB biofilm formation	biofilm formation	+/+	-/-	-/-	-/-
302	hmsH, pgaA, biofilm formation	biofilm formation	+/+	-/-	-/-	-/-
1963	hmsP negative regulator of biofilm formation in y. pestis	biofilm formation	+/+	-/-	+/+	-/-
2010	hmsR/icaA IM protein, polymerizes ?-1,6-GlcNAc, biofilm PIA synthesis N-glycosyltransferase icaA (EC	biofilm formation	+/+	-/-	-/-	-/-
2275	hmsS, pgaD, biofilm formation Y.pestis	biofilm formation	+/+	-/-	-/-	-/-
300	icaB, notB, cda, polysaccharide or chitin deacetylase (icaB in icaADBC)	biofilm formation	+/+	-/-	+/-	+/-
2015	icaC, biofilm PIA synthesis protein icaC, biofilm formation along with icaABC	biofilm formation	-/-	+/-	-/-	-/-
2013	icaD, biofilm PIA synthesis protein icaD, biofilm formation along with icaABC	biofilm formation	-/0	-/-	-/-	-/-
2429	gmhA putative sedoheptulose 7-phosphate isomerase, phosphoheptose isomerase (EC 5.3.1.-)	capsular polysaccharide synthesis	+/+	-/-	+/+	-/-
2457	manC, putative GDP-mannose pyrophosphorylase, mannose-1-phosphate guanylyltransferase [GDP] (EC 2.7.	capsular polysaccharide synthesis	+/+	-/-	+/+	-/-
2455	wcbA, lipA, bexD, putative capsular polysaccharide export protein, polysaccharide export outer membr	capsular polysaccharide synthesis	-/-	-/-	-/-	-/-
2450	wcbD, bexC, ctrB, putative capsular polysaccharide ABC transporter, inner-membrane transmembrane pro	capsular polysaccharide synthesis	-/-	-/-	-/-	-/-
2433	wcbK GDP sugar epimerase/dehydratase	capsular polysaccharide synthesis	+/+	-/-	+/+	-/-
2425	wcbN putative D-glycero-d-manno-heptose 1,7-bisphosphate phosphatase, (EC 3.1.3.-)	capsular polysaccharide synthesis	+/+	-/-	+/-	-/-
2423	wcbO capsule polysaccharide biosynthesis and export protein	capsular polysaccharide synthesis	-/-	-/-	-/-	-/-
2421	wcbP dehydrogenase/reductase protein	capsular polysaccharide synthesis	+/+	+/-	+/+	-/-
2417	wcbR type I polyketide synthase, fatty acid synthase (EC 2.3.1.-)	capsular polysaccharide synthesis	-/-	-/-	-/-	-/-
2415	wcbS UDP-3-O-[3-hydroxymyristoyl] N-acetylglucoseamine deacetylase, UDP-3-O-acyl-GlcNAc deacetylase.	capsular polysaccharide synthesis	+/+	-/-	+/+	-/-
2413	wcbT acyl-CoA transferase (EC 2.3.1.-)	capsular polysaccharide synthesis	+/+	+/+	+/+	-/-
2446	wzt2, bexA, putative ABC transporter for capsular polysaccharide (teichoic acids export tagH, capsul	capsular polysaccharide synthesis	-/-	+/0	+/-	+/-
8583	capD, required for capsule biosynthesis in Bacillus anthracis	capsule biosynthesis	+/+	+/-	+/+	-/-
2342	bicA, sycD, ipgC, lcrH, sicA, ygeG type III secretion chaperone	chaperone for type III secretion	+/+	-/-	+/-	-/-
114	cheW, chemosensation adapter (H.pylori cheW)	chemotaxis	+/+	-/-	+/-	-/-
112	cheY like two component response regulator	chemotaxis	+/+	+/+	+/+	+/-
218	fla (A,B,C,E), flagellin	flagella	-/+	-/-	-/-	-/-
7221	fla(A,B), fli(C,D,J,K,M,N,O), filament protein flagella	flagella	+/+	-/-	+/-	-/-

Supplementary Table 1

237	flgC, flagellar basal-body protein	flagella	+/+	-/-	+/-	-/-
227	flgD, flagellar hook assembly protein, hook capping protein	flagella	+/+	-/-	+/-	-/-
221	flgE, flagellar hook protein	flagella	-/+	-/-	-/+	-/-
7234	flgE, flagellar hook protein	flagella	+/+	-/-	+/+	-/-
241	flgG, flagellar basal-body protein	flagella	+/+	-/-	+/+	-/-
229	flgH, flagellar basal-body L-ring protein	flagella	+/+	-/-	+/+	-/-
233	flgI, flagellar basal-body P-ring protein	flagella	+/+	-/-	+/-	-/-
7310	flgK, between flgL and flgE, hook associated protein 1	flagella	+/+	-/-	+/+	-/-
223	flgK, flagellar hook associated protein (hap1)	flagella	+/+	-/-	+/+	-/-
225	fliD, flagellar cap protein, possible adhesin (hap2)	flagella	+/+	-/-	+/-	-/-
235	fliE, flagellar basal-body protein	flagella	+/+	-/-	+/0	-/-
231	fliF, flagellar basal-body M-ring protein	flagella	+/+	-/-	+/+	-/-
247	fliG, flagellar motor switch protein	flagella	+/+	-/-	+/+	-/-
261	fliI, flagellar export protein	flagella	+/+	-/-	+/+	-/-
251	fliM, flagellar motor switch protein	flagella	+/+	-/-	+/+	-/-
249	fliN, flagellar switch protein	flagella	+/+	-/-	+/-	-/-
270	fliR, flagellar biosynthesis protein	flagella	+/+	-/-	+/-	-/-
255	fliS, flagellin-specific chaperone	flagella	+/+	-/-	+/-	-/-
2243	motA flagellar motor rotation, proton motor (H.pylori motA)	flagella	+/+	-/-	+/-	-/-
7190	motB, along with motA, converts proton energy into torque for flagellar motor	flagella	+/+	-/-	+/-	-/-
506	secA, secA2 general secretory pathway	GSP	+/+	+/+	+/+	+/0
512	secE, general secretory pathway	GSP	+/+	+/-	+/+	-/-
514	secG, general secretory pathway	GSP	+/+	+/0	+/+	+/-
509	secY, general secretory pathway	GSP	+/+	+/+	+/+	+/-
7270	2,3-dihydro-2,3-dihydroxybenzoate dehydrogenase	iron acquisition-siderophore synthesis	+/+	-/-	+/+	-/-
2530	BPSS1775, Ferrichrysobactin, Ferrioxamine, Fe(III)-pyochelin receptor, possible tonB dependent recep	iron acquisition-siderophore synthesis	+/+	-/-	+/+	-/-
2532	BPSS1776, flavin binding monooxygenase, hydroxamate siderophore synthesis	iron acquisition-siderophore synthesis	-/-	-/-	-/-	-/-
2534	BPSS1777, putative siderophore related non-ribosomal peptide synthetase	iron acquisition-siderophore synthesis	+/+	-/-	+/-	-/-
2536	BPSS1778, putative siderophore related non-ribosomal peptide synthetase	iron acquisition-siderophore synthesis	-/+	-/-	-/+	-/-
7268	entB like, isochorsmatase, siderophore biosynthesis, enterochelin synthase B (EC 3.3.2.1)	iron acquisition-siderophore synthesis	+/+	-/-	+/+	-/-
7211	entD, enterobactin synthetase component D, 4-phosphopantetheinyl transferase entD (EC 2.7.8.-)	iron acquisition-siderophore synthesis	+/+	-/-	+/-	-/-
313	irp1 (HMWP1), yersiniabactin synthesis	iron acquisition-siderophore synthesis	-/-	-/-	-/-	-/-
2544	irp5, ybtE, pchD salicyl-AMP ligase, salicylate-activating enzyme, siderophore biosynthesis, enterob	iron acquisition-siderophore synthesis	+/+	-/-	+/+	-/-
2538	pchA, entC, dhbC, putative salicylate biosynthesis isochorismate synthase,(EC 5.4.4.2) or anthranil	iron acquisition-siderophore synthesis	+/+	+/-	+/0	-/-
2542	pchC, lipase similar to pyochelin biosynthetic protein from Pseudomonas aeruginosa	iron acquisition-siderophore synthesis	-/-	-/-	-/0	-/-
2548	pchE, irp2, HMWP2 pyochelin/yersiniabactin synthetase	iron acquisition-siderophore synthesis	-/-	-/-	-/-	-/-
2550	pchF, putative siderophore related non-ribosomal peptide synthetase	iron acquisition-siderophore synthesis	-/-	-/-	-/+	-/-
54	siderophore synthesis lucA/C (F.tularensis/B.anthracis frgA, lucA, lucC)	iron acquisition-siderophore synthesis	-/+	+/-	-/-	-/-
372	ybtQ (irp7) siderophore transport	iron acquisition-siderophore synthesis	-/-	-/-	-/-	-/-

Supplementary Table 1

374	ybtS (irp9) siderophore biosynthesis	iron acquisition-siderophore synthesis	-/-	-/-	-/-	-/-
376	ybtT (irp4) siderophore biosynthesis	iron acquisition-siderophore synthesis	-/-	-/-	-/-	-/-
378	ybtU/pchG siderophore biosynthesis, reductase	iron acquisition-siderophore synthesis	-/-	-/-	-/-	-/-
160	fecA, iron(III) dicitrate transporter (H.pylori fecA)	iron acquisition-siderophore transport	-/-	-/-	-/-	-/-
296	fec(C,D) family (eg. hmuU [hmuTUV] in Y.pestis) iron transport IM permease, ABC-type Fe3+-siderophor	iron acquisition-siderophore transport	+/+	+/+	+/+	+/0
7256	fecE like iron(III) siderophore, dicitrate transport ATP binding protein (fecE like)	iron acquisition-siderophore transport	+/+	+/0	+/+	+/-
8562	fhuB, duplicated permease domains	iron acquisition-siderophore transport	+/+	+/+	+/+	+/-
2248	fhuC, ABC-type cobalamin/Fe3+-seiderophores transport systems, ATPase component	iron acquisition-siderophore transport	+/+	+/0	+/-	+/-
286	fhuD, ferrichrome-binding periplasmic protein precursor	iron acquisition-siderophore transport	+/+	-/-	+/-	-/-
288	fhuF, (ysuF) ferric iron reductase	iron acquisition-siderophore transport	+/+	-/-	+/-	-/-
292	hmuR (hemR), a TonB dependent OM receptor for hemoproteins	iron acquisition-siderophore transport	+/+	-/0	-/-	-/-
290	hmuS, (hemS) hemin degredation	iron acquisition-siderophore transport	+/+	-/-	-/-	-/-
294	hmuT, siderophore binding protein, part of an ABC transporter (hmuTUV)	iron acquisition-siderophore transport	+/+	-/-	-/-	-/-
298	hmuV, ATP binding protein, part of an ABC transporter (hmuTUV)	iron acquisition-siderophore transport	+/+	+/-	+/+	+/-
342	psn, OM ligand gated channel, yersiniabactin receptor, pesticin receptor	iron acquisition-siderophore transport	-/-	-/-	-/-	-/-
370	ybtP (irp6) siderophore transport	iron acquisition-siderophore transport	-/-	-/-	-/-	-/-
44	major iron response regulator(F.tularensis/H.pylori Fur) along with iron, acts to repress transcript	iron acquisition-transcriptional regulation	+/+	+/+	+/+	+/-
170	ceuE, shuttles Fe(III) in the periplasm (H.pylori ceuE)	iron acquisition-transport	-/-	-/-	+/-	-/-
7260	fbpC like, iron(III) transporter ATP-binding protein (sfuC like)	iron acquisition-transport	+/+	-/-	-/-	-/-
173	feoB, inner membrane Fe(II)transport protein (H.pylori feoB)	iron acquisition-transport	+/+	-/-	+/+	-/-
7262	sfuC or cysA iron(III) transport ATP binding protein (sfuC like) or Sulfate/thiosulfate import ATP	iron acquisition-transport	+/+	+/+	+/+	+/-
380	yfeA, periplasmic binding protein	iron acquisition-transport	-/-	-/-	+/-	-/-
1977	yfe(C,D) ABC-type Fe/Mn/Zn transporter, permease component	iron acquisition-transport	-/-	+/+	+/0	+/0
52	tppB tri-peptide permease, possible di- or tri- peptide transporter (F.tularensis iraB)	iron acquisition-transport peptide	+/+	-/-	+/+	-/-
2485	apaH, bis(5-nucleosyl)-tetraphosphatase (EC 3.6.1.41)	LPS biosynthesis	+/+	-/-	+/+	-/-
2483	rmlB, spsJ, dTDP-glucose 4,6 dehydratase, spore coat polysaccharide biosynthesis protein (EC 4.2.1.	LPS biosynthesis	+/+	-/-	+/+	+/-
2472	wbiB, epimerase/dehydratase. putative UDP-glucose 4-epimerase (EC 5.1.3.2)	LPS biosynthesis	-/+	-/-	-/-	-/-
2465	wbi(E,F) glycosyl transferase (EC 2.-.-.-)	LPS biosynthesis	+/+	+/+	+/+	+/-
2461	wbiH, undecaprenyl phosphate N-acetylglucoseaminyltransferase, UDP-GlcNAc:undecaprenyl-P GlcNAc 1-P	LPS biosynthesis	-/-	+/-	-/-	+/-
2459	wbiI, capD epimerase/dehydratase	LPS biosynthesis	-/+	+/+	-/-	-/-
2476	wzt, ABC transporter, ATP binding component	LPS biosynthesis	-/-	+/-	-/-	+/-
2	iglA like, unknown activity but required for virulence, gene allows survival within the macrophage (	macrophage colonization	+/-	-/-	+/-	-/-
9	macrophage infectivity potentator like protein(F.tularensis FTT1043)	macrophage colonization	+/+	-/-	+/+	-/-
60	minD, possible pump for toxic or radical ions(F.tularensis minD)	macrophage colonization	+/+	-/-	+/+	-/-

11	M. tuberculosis mce homolog (F.tularensis FTT1249)	macrophage colonization	+/+	-/-	+/+	-/-
183	phoP, response regulator	macrophage colonization	+/+	-/-	+/-	-/-
1973	PhoQ, histadine Kinase	macrophage colonization	+/+	-/-	+/-	-/-
56	sspA, stringent starvation protein A, global transcriptional regulation(F.tularensis mglA)	macrophage colonization	+/+	-/-	+/+	-/-
58	sspB, stringent starvation protein B, global transcriptional regulation(F.tularensis mglB)	macrophage colonization	+/+	-/-	+/+	-/-
8389	AbgT, efflux pump component, putative transporter family	multi-drug efflux pump	+/+	+/+	-/-	-/-
2566	acrA/ttg(A,D,G)/mdtE like putative mulit-drug efflux protein	multi-drug efflux pump	+/+	-/-	+/+	-/-
8333	AcrB/AcrD/AcrF family, cation/multidrug efflux pump	multi-drug efflux pump	+/+	-/-	+/+	-/-
2568	acr(B,D,F)/mdt(B,C,F)/ttg(B,E,H) multi-drug efflux protein	multi-drug efflux pump	+/+	-/-	+/+	-/-
8385	agrA like, aminoglycoside resistance efflux transporter	multi-drug efflux pump	+/+	-/-	+/+	-/-
8413	arpC, mepC, opr(J,M), sepC, srpC, ttg(C,F,I) outer membrane efflux pump	multi-drug efflux pump	+/+	-/-	+/-	-/-
8535	ebr (A,B), emrE, mdtJ, qac(E,F) multi-drug resistance	multi-drug efflux pump	+/+	-/-	+/-	-/-
8453	emr(A,K) multi-drug resistance protein	multi-drug efflux pump	+/+	-/-	+/+	-/-
7286	emr(B,Y) multi-drug resistance protein	multi-drug efflux pump	+/+	-/+	+/+	-/0
8479	lmrA multi-drug resistance ABC transporter ATP-binding and permease protein	multi-drug efflux pump	-/-	+/+	-/-	-/-
2586	macA macrolide-specific multi-drug efflux pump	multi-drug efflux pump	+/+	-/-	+/-	-/-
2584	macB macrolide-specific ABC type efflux carrier	multi-drug efflux pump	+/+	-/-	+/-	+/0
8313	major facilitator superfamily export protein	multi-drug efflux pump	+/+	+/-	+/+	-/-
8298	major facilitator superfamily protein, ydiC	multi-drug efflux pump	+/+	-/+	+/-	-/-
8286	mdtA multi-drug transporter	multi-drug efflux pump	+/+	-/-	+/+	-/-
8363	mdtE like, periplasmic protein (HlyD family secretion protein)	multi-drug efflux pump	+/+	-/-	+/+	-/-
8355	mdtG, pmrA major-facilitator superfamily	multi-drug efflux pump	+/+	-/-	+/-	+/-
8469	mdtK, norM multi-drug resistance protein Na(+)/drug antiporter	multi-drug efflux pump	+/+	-/-	+/-	-/-
8254	norA multi-drug resistance protein	multi-drug efflux pump	-/-	+/+	+/-	+/-
2579	norM, multi-drug resistance protein	multi-drug efflux pump	+/+	-/-	+/-	-/-
8268	putative antibiotic efflux protein (transporter)	multi-drug efflux pump	+/+	+/-	+/-	+/-
8353	putative efflux pump	multi-drug efflux pump	+/+	-/-	+/-	-/-
8315	putative major facilitator superfamily export protein	multi-drug efflux pump	+/+	-/-	+/0	-/-
8529	ykkC multi-drug resistance protein	multi-drug efflux pump	-/-	+/-	-/-	-/-
8531	ykkD multi-drug resistance protein	multi-drug efflux pump	-/-	+/-	-/-	-/-
148	gdh, glutamate dehydrogenase (EC 1.4.1.2) (H.pylori gdhA)	nitrogen metabolism	+/+	+/+	+/+	-/-
146	glutamine synthetase, Glutamate--ammonia ligase, (EC 6.3.1.2) (H.pylori glnA)	nitrogen metabolism	+/+	+/+	+/+	+/-
363	ure1, ureC, allpha subunit of the urease enzyme, urea amidohydrolase (EC 3.5.1.5)	nitrogen metabolism-urease	+/+	+/+	-/-	-/-
361	ure2, ureB, beta subunit of the urease enzyme, urea amidohydrolase (EC 3.5.1.5)	nitrogen metabolism-urease	+/+	+/-	-/-	-/-
359	ure3, ureA, gamma subunit of the urease enzyme, urea amidohydrolase (EC 3.5.1.5)	nitrogen metabolism-urease	+/0	+/-	-/-	-/-
126	ureA, Alpha subunit of urease holoenzyme (H.pylori ureB)	nitrogen metabolism-urease	+/0	+/+	-/-	-/-
124	ureB, Beta/Gamma subunit of urease holoenzyme (H.pylori ureA)	nitrogen metabolism-urease	+/0	+/-	-/-	-/-
128	ureE, urease assembly protein (H.pylori ureE)	nitrogen metabolism-urease	-/0	+/-	-/-	-/-
130	ureF, urease assembly protein (H.pylori ureF)	nitrogen metabolism-urease	+/0	+/+	-/-	-/-
132	ureG, urease assembly protein (H.pylori ureG)	nitrogen metabolism-urease	+/0	+/+	-/-	-/-
134	ureH, urease assembly protein (H.pylori ureH)	nitrogen metabolism-urease	+/0	+/-	-/-	-/-
1959	nixA, nic1 high-affinity Nickel transporter	nitrogen metabolism-urease activity	-/0	-/-	+/-	-/-

Supplementary Table 1

1957	ureI Urea transporter	nitrogen metabolism-urea transport	-/+	+/-	-/-	-/-
2518	hmpA, NOD, flavohaemoprotein, Nitric oxide dioxygenase, NO oxygenase(EC 1.14.12.17)	NO resistance	+/+	+/-	+/0	-/-
2515	catE, catA, catB, catylase HP II, (EC 1.11.1.6)	peroxide resistance	+/+	+/+	+/+	-/-
317	katY catalase-peroxidase	peroxide resistance	+/+	-/-	+/+	-/-
2508	oxyR or estR, oxidative stress regulatory protein, hydrogen peroxide-inducible genes activator or es	peroxide resistance	+/+	-/-	+/+	-/-
336	psaB, myfB, caf1M, psaB, papD, etc. fimbrial chaperone	pH 6 antigen pili	+/+	-/-	+/0	-/-
1955	psaC, papC, aggC, fimC, fimD, etc. Usher protein, outer memebrane protein	pH 6 antigen pili	+/+	-/0	+/0	-/-
340	psaF, pH 6 antigen biosynthesis	pH 6 antigen pili	-/0	+/0	-/-	-/-
329	plasminogen activator, coagulase/fibrinolysin precursor (EC 3.4.23.48) or protease VII precursor (EC	plasminogen actitvator	+/+	-/-	+/-	-/-
2495	probable serine protease do-like precursor (EC 3.4.21.-)	protease	+/+	+/-	+/+	+/-
8589	agrB like, integral membrane protein that modifies agrD (AIP), processing it to become active. Part	quorum sensing	-/-	+/-	-/-	-/-
526	rmlA, rhamnose/LPS biosynthesis, glucose-1-phosphate thymidylyltransferase (EC 2.7.7.24) (EC 2.7.7.2	rhamnose synthesis	+/+	-/-	+/-	+/-
530	rmlB, acbB, rhamnose biosynthesis, dTDP-glucose 4,6-dehydratase (EC 4.2.1.46)	rhamnose synthesis	+/+	-/-	+/+	+/-
528	rmlC, rfbC rhamnose/LPS biosynthesis, dTDP-4-dehydrorhamnose 3,5-epimerase (EC 5.1.3.13)	rhamnose synthesis	-/-	-/-	+/-	-/-
532	rmlD, rfbD, rhamnose/LPS biosynthesis, dTDP-4-dehydrorhamnose reductase (EC 1.1.1.133)	rhamnose synthesis	-/-	-/-	+/-	+/+
42	Carbohydrate/purine kinase, pfkB family(F.tularensis FTT0801c)	surface antigen synthesis	+/+	+/-	+/+	+/-
37	glycosyltransferase 2(F.tularensis FTT0797, FTT0798)	surface antigen synthesis	-/+	-/0	-/-	-/-
23	probable sugar transferase (F.tularensis FTT0790)	surface antigen synthesis	+/+	+/-	+/-	-/-
21	PRTaseII phosphoribosyl transferase (F.tularensis FTT0789), Ribulose-phosphate 3-epimerase(EC 5.1.3.	surface antigen synthesis	+/+	+/-	+/+	+/-
25	putative UDP-glucose 4-epimerase(F.tularensis FTT0791)	surface antigen synthesis	+/+	+/-	+/+	-/-
93	ahpC, peroxiredoxin, alkyl hydorperoxide reductase, thioredoxin peroxidase (EC 1.11.1.15) (H.pylori	survive oxidative stress	+/+	+/+	+/+	+/-
91	ahpF (EC 1.6.4.-) or thioredoxin reductase (EC 1.8.1.9) (H.pylori trxB)	survive oxidative stress	+/+	+/+	+/+	+/-
140	arginase (EC 3.5.3.1) (H.pylori rocF)	survive oxidative stress	-/-	+/+	-/-	-/-
95	bcp, putative peroxiredoxin, Thioredoxin reductase, (EC 1.11.1.15) (H.pylori bcp)	survive oxidative stress	+/+	+/-	+/+	-/-
2513	cat(A,B) catylase precursor, has signal peptide, (EC 1.11.1.6)	survive oxidative stress	+/+	+/+	+/+	-/-
50	Fe/Mn regulated superoxide dismutase (EC 1.15.1.1) sodB, sodF, sodM (F.tularensis sodB)	survive oxidative stress	+/+	+/+	+/+	+/0
46	ferritin, bacterioferritin (F.tularensis/H.pylori ftnA, pfr, napA)	survive oxidative stress	+/+	+/+	+/+	+/-
144	nifS, synthesis of [Fe-S] cluster (H.pylori nifS)	survive oxidative stress	+/+	+/-	+/+	-/-
2246	peroxiredoxin (tpx) thiol peroxidase (EC 1.11.1.-)	survive oxidative stress	+/+	+/-	+/-	-/-
2511	sodC, putative membrane attached superoxide dismutase, Cu-Zn, (EC 1.15.1.1)	survive oxidative stress	+/+	-/-	-/-	-/-
89	trx(1,2,3,F, H,M) thioredoxin and thioredoxin-like proteins (H.pylori trxA)	survive oxidative stress	+/+	+/+	+/+	+/+
2616	cpaA, pilus assembly, prepillin peptidase, type 4 prepilin-like proteins leader peptide processing e	tad-type pilus	+/-	-/0	+/-	-/-
7326	tat(A,E), TAT secretion, sec independent protein translocase	TAT secretion	+/+	-/-	+/+	-/-
7324	tatC, TAT secretion, sec independent protein translocase	TAT secretion	+/+	-/-	+/+	-/-
7330	tatD, TAT secretion, sec independent protein translocase, deoxyribonuclease	TAT secretion	+/+	+/-	+/+	+/-
2144	ccdA anti-toxin to ccdB	toxin	+/+	-/-	-/-	-/-
2104	Cl- channel inhibitor, neurotoxin	toxin	-/-	-/-	-/-	-/-
2098	delta lysin Staphylococcus aureus	toxin	-/-	-/-	-/-	-/-
1982	EAST1 (AstA) Enteroaggregative heat stable enterotoxin	toxin	+/-	-/-	+/-	-/-
2156	entericidin(A,B), antitoxin/toxin pair	toxin	+/+	-/-	+/-	-/-
1990	esp enterococcal surface protein, Enterococcus faecium	toxin	-/+	-/-	-/+	-/-

Supplementary Table 1

1992	gelE coccolysin protease, Enterococcus faecium also thermolysin, sepp1, bacillolysin, pseudolysin, a	toxin	-/-	+/-	-/-	-/-
7205	hemolysin III	toxin	+/+	+/-	+/-	-/-
2004	hlyC, (hlyA activating enzyme), hemolysin/cytolysin/leukotoxin/RTX-activating lysine-acyltransferase	toxin	+/-	-/-	-/-	-/-
2118	hlyE pore forming toxin of E.coli	toxin	+/+	-/-	+/+	-/-
2000	hylA, RTX-(I,II,III), leukotoxin, cytolysin, hemolysin, protoxin	toxin	+/-	-/-	-/-	-/-
2086	icsA Shigella flexneri actin polymerization	toxin	+/+	-/-	+/-	-/-
2102	Mu conotoxins neurotoxin	toxin	+/+	-/-	+/+	+/-
1986	one or two component bacterial exotoxin with membrane-damaging function (cytK B.cereus, lukNS S.aure	toxin	-/-	-/-	-/-	-/-
2199	pemK toxin, cleaves single stranded RNA, E.coli plasmid pC15-1a	toxin	+/+	+/+	-/-	-/-
2022	SEA, SPEA, scarlet fever toxin, enterotoxin (A,B,C,D,E,G,H), exotoxin(A,C,G,H) superantigen S. a	toxin	-/-	-/-	-/-	+/-
274	tlyA, pore-forming hemolysin	toxin	-/-	-/-	-/-	+/-
272	tlyC, possible hemolysin	toxin	+/+	+/-	+/+	-/-
2024	TST, TSST-1 superantigen responsible for toxic-shock syndrome	toxin	-/-	-/-	-/-	-/-
2128	zeta bacterial toxin, part of toxin/anti-toxin pair, from a plasmid	toxin	-/-	-/-	-/-	-/-
502	prsA, aids folding of secreted proteins B. anthracis	toxin folding	-/-	+/-	-/-	-/-
2354	bapC, iagB, ipgF like, type III secretion effector protein	toxin-type III secretion	+/+	-/-	+/-	-/-
2112	agrA, response regulator for quorum-sensing	transcriptional control	+/+	+/-	+/-	+/-
2110	agrC, quorum-sensing sensor kinase	transcriptional control	-/-	+/-	-/-	-/-
2108	agrD, peptide activator of quorum-sensing agrC sensor kinase	transcriptional control	-/-	+/-	-/-	-/-
308	ihfAB, integration host factor or DNA-binding protein HU-(alpha,beta)	transcriptional control	+/+	+/+	+/+	+/+
382	ymoA, hha, thermoregulation of virF (regulator of the yop virulon), haemolysin expression-modulating	transcriptional control	+/+	-/-	+/-	-/-
500	acpA, acpB, atxA B. anthracis transcriptional activators	transcriptional regulation	-/-	-/-	-/-	+/-
2362	bprB like, type III secretion regulation, response regulator	transcriptional regulation	+/+	-/-	+/+	-/-
166	exbB, tolQ, energy providing protein required for some transporter or flagellar function (H.pylori	transporter	+/+	-/-	+/+	-/-
2245	exbD energy providing (H.pylori)	transporter	+/+	-/-	+/-	-/-
185	exbD, tolR, produces energy for some transporters and flagella (H.pylori exbD)	transporter	+/+	-/-	+/+	-/-
2132	tolA, colicin import protein	transporter	+/+	-/-	+/+	-/-
164	tonB, energy providing protein required for some transporter function (H.pylori TonB)	transporter	+/+	-/-	+/+	-/-
2594	fimA, major or minor fimbrial subunit	type I fimbriae	+/+	-/-	+/-	-/-
2590	papC, usher protein, P pilus assembly protein	type I fimbriae	+/+	-/-	+/+	-/-
2592	papD, fimbrial chaperone, P pilus assembly protein	type I fimbriae	+/+	-/-	+/-	-/-
2322	bsaQ, lcrD, invA, virH like, type III secretion system protein	type III secretion	+/+	-/-	+/+	-/-
2326	bsaS, invC, spaL, mxiB, ssaN like, type III secretion system protein, similar to ATP synthase beta c	type III secretion	+/+	-/-	+/+	-/-
2334	bsaW, hrcR, spaP, yscR like, type III secretion system protein	type III secretion	+/+	-/-	+/-	-/-
2336	bsaX, spaQ, yscS like, type III secretion system protein	type III secretion	+/+	-/-	+/-	-/-
430	yscN, hrcN, hrpB6, type III secretion ATP synthase	type III secretion	+/+	-/-	+/+	-/-
442	yscT, type III secretion	type III secretion	+/+	-/-	+/-	-/-
2340	yscU, bsaZ, sctU, hrcU, spaS, ssaU type III secretion protein	type III secretion	+/+	-/-	+/+	-/-
422	yscJ, hrpB3, ssaJ or FliF or nolT	type III secretion or flagella	+/+	-/-	+/+	-/-
438	yscR, hrcR, hrpW or fliP, type III secretion or flagellar component	type III secretion or flagella	+/+	-/-	+/-	-/-
263	YscR or FliP	type III secretion or flagella	+/+	-/-	+/-	-/-
268	YscS or FliQ	type III secretion or flagella	+/+	-/-	+/-	-/-

440	yscS, ssaS or fliQ, component of type III secretion system or Flagella (FliQ) (Y.pestis yscS)	type III secretion or flagella	+/+	-/-	+/-	-/-
239	YscU or FlhB	type III secretion or flagella	+/+	-/-	+/+	-/-
446	yscV, hrpI, hrcV, hrpC2, ssaV or flhA, component of type III secretion system or Flagella (FlhA) (Y.	type III secretion or flagella	+/+	-/-	+/+	-/-
266	YscV or FlhA	type III secretion or flagella	+/+	-/-	+/+	-/-
2075	hlyB, rtxB, cyaB, lktB (VC1448) type I secretion channel, ATP binding IM protein (transport RTX, leu	type I secretion	+/+	+/+	-/+	-/-
2077	hlyD, rtxD, cyaD, lktD type I secretion	type I secretion	+/-	-/-	+/-	-/-
2008	tolC outer membrane channel, part of type I secretion system, RND and MFS multi-drug-efflux systems.	type I secretion	+/+	-/-	+/+	-/-
2695	pilT, twitching mobility protein, periplasmic location, promotes disassembly of pilus	type IVB pili	+/+	-/-	+/+	-/-
70	pilT, required for pilus retraction/depolymerization (F.tularensis pilT)	type IV pili	+/+	-/-	+/+	-/-
74	pilus structural unit, minor pilin(F.tularensis pilE5, pilE4)	type IV pili	+/+	-/-	+/0	-/-
68	pppA, prepilin peptidase, required to produce functional pilin(F.tularensis pilD)	type IV pili	+/-	+/-	-/-	-/-
80	hofB type IV pilus protein or gspE, ATPase, substrate transfer, unfolding, chaperone type IV pilus o	type IV pilus assembly or type II secretion	+/+	+/-	+/+	+/-
82	hofC type IV pilus protein, flaJ or gspF, inner membrane protein involved in type IV pilus assembly	type IV pilus assembly or type II secretion	+/+	-/-	+/+	+/-
64	pilQ, gspD, hofQ outer membrane pore, gated channel, required for pilus assembly similar to gspD (F	type IV pilus assembly or type II secretion	+/+	-/-	+/0	-/-
194	virD4, traG, type IV secretion protein(H.pylori cag5)	type IV secretion	-/-	-/-	-/-	-/-
589	DUS1, H1L vaccinia virus dual specificity protein phosphatase (EC 3.1.3.48) (EC 3.1.3.16)	viral effector	+/+	-/-	-/-	-/-

Species	**Strain**	**GenBank Accession**
Escherichia coli	CFT073	AE014075.1
Escherichia coli	K-12 MG1655	U00096.2
Enterococcus faecalis	V583	AE016830.1
Staphylococcus aureus subsp. *aureus*	Mu50	NC_002758.2

Table 1: Bacterial Strains Used for Probe Selection

Primer3 parameter	**Value**
PRIMER_TASK	pick_hyb_probe_only
PRIMER_PICK_ANYWAY	1
PRIMER_INTERNAL_OLIGO_OPT_SIZE	60
PRIMER_INTERNAL_OLIGO_MIN_SIZE	50
PRIMER_INTERNAL_OLIGO_MAX_SIZE	66
PRIMER_INTERNAL_OLIGO_OPT_TM	90
PRIMER_INTERNAL_OLIGO_MIN_TM	80
PRIMER_INTERNAL_OLIGO_MAX_TM	150
PRIMER_INTERNAL_OLIGO_MIN_GC	25
PRIMER_INTERNAL_OLIGO_MAX_GC	75
PRIMER_EXPLAIN_FLAG	0
PRIMER_INTERNAL_OLIGO_SALT_CONC	450
PRIMER_INTERNAL_OLIGO_DNA_CONC	100
PRIMER_INTERNAL_OLIGO_MAX_POLY_X	4
Other parameters	defaults

Table 2: Input parameters for Primer3 probe generation