

Leveraging Complete Genome Sequences in the Design of a Compact Multi-organism *E. coli* mRNA Expression MicroArray.

Christopher Davies¹, Gangwu Mei², Brant Wong, Harley Gorrell, Alan Williams

Keywords: Escherichia coli, microarray, gene expression, Affymetrix, mRNA

1 Introduction.

Since the completion of the first genomic sequence for *Escherichia coli* in 1997 (GenBank accession no. NC_000913) [1], upon which Affymetrix based its version 1 oligonucleotide microarray to measure gene expression in *E. coli* [2], there have been 3 more *E. coli* complete genomic sequences released into the public domain (GenBank accession no. NC_002655, NC_002695, NC_004431) [3, 4, 5]. These 3 are for enteropathogenic and uropathogenic strains of this organism, and present a serious public health risk. Several deaths have occurred due to food contamination by two of these strains.

Affymetrix has designed a multi-organism gene expression array that takes advantage of the highly homologous (95% identity) set of core genes in these organisms. The array contains 25-mer probe sets that target the identical, shared, regions in these orthologous genes, whilst avoiding the 5% polymorphisms (one in every 20 bases) that would make it more of a challenge to design an array based on long-mer oligos (e.g 50 and 70+ mers). This compression of the genomic content means that the core genes and the unique genes that distinguish each pathogenic strain, all fit on an array that is no larger than the v1 array.

2 Design Strategy.

Given the curation that has occurred on the original MG1655 genomic sequence, Affymetrix decided that a revised *E. coli* genome array was called for, in addition one that would allow the measurement of gene expression in the pathogenic strains. The 3 pathogenic *E. coli* genomes are all larger than the laboratory strain, K12:MG1655, each containing approximately 1000 more genes than MG1655 (which has approx. 4358 protein coding genes). Thus, an array to measure expression in all these organisms might contain probes specific for approx 22,000 genes, which would make it considerably larger than the v1 array.

We wished to avoid increasing the size of the array. It is known that a core set of conserved genes exists in these organisms ([5]). We postulated that if the DNA sequence similarity was great enough between these core genes, a single probe set might suffice to measure gene expression for a family of orthologs present in more than one of the organisms. Preliminary blast analysis showed that this

¹ Affymetrix Corp., 6550 Vallejo, Emeryville, CA 94608 E-mail: Christopher_Davies@Affymetrix.com

² Affymetrix Corp., 6550 Vallejo, Emeryville, CA 94608 E-mail: Gangwu_Mei@Affymetrix.com

was indeed the case. A core set of ~3000 genes was found to be more than 95% identical over their entire length in all 4 strains, (whereas at 99% identity, only 200-500 genes are identical over their length). With this information in hand, and by making use of the shared probe set strategy, we were able to create an array that was the same size as the V1 array, containing 10,014 probe sets that measure the expression of 20,366 genes across all 4 strains (Table 1).

Strain	<i>Genome gene counts</i>	<i>Tiled probe sets</i>	<i>Probe set counts</i>	<i>Represented gene counts</i>
MG1655	4,358	4,070	5,298	4,358
EDL933	5,376	1,787	5,480	5,376
CFT073	5,379	2,486	7,421	5,379
SAKAI	5,253	373	5,380	5,253
Intergenic	4,534	1,298	1,427	1,427
totals	25,900	10,014		

Table 1: Number of genes, probe sets' origin, number of probe sets that measure each strain, and the numbers of genes represented by those probe sets for each strain.

References

- [1] Blattner et al, 1997. The complete genome sequence of Escherichia coli K-12. *Science*. 1997 Sep 5;277(5331):1453-74.
- [2] <http://www.affymetrix.com/support/technical/byproduct.affx?product=ecolisense>.
- [3] Perna et al, 2001. Genome sequence of enterohaemorrhagic Escherichia coli O157:H7. *Nature*. 2001 Jan 25;409(6819):529-33.
- [4] Hayashi et al, 2001. Complete genome sequence of enterohemorrhagic Escherichia coli O157:H7 and genomic comparison with a laboratory strain K-12. *DNA Research* 2001 Feb 28;8(1):11-22.
- [5] Welch et al, 2002. Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli. *Proceedings of the National Academy of Sciences USA* . 2002 Dec 24;99(26):17020-4.