UCL DEPARTMENT OF
COMPUTER SCIENCE

Research Note

RN/12/11

Correlation of Microarray Probes give Evidence for Mycoplasma Contamination in Human Studies

2 November 2012

W. B. Langdon

Abstract

Recently the complete genomes of several more species of mycoplasma have been published. We can now see at least 473 Affymetrix HG-U133 +2 Homosapiens probes match one or more species of mycoplasma. Analysis of published data from thousands of human GeneChips finds correlations between probes in human studies between different labs in different countries which suggests contamination with mycoplasma is the common factor.

1 Introduction

It is well known that mycoplasma contamination is rife in molecular biology laboratories [6]. Depending upon medium, mycoplasma contamination rates of 1% to 15–35% (or even higher) have been reported [7]. Many labs routinely sterilised their equipment to counter it. About 1% of published NCBI's GEO GeneChip data appear to be contaminated [8]. Aldecoa-Otalora *et al.* [8] showed one complete Affymetrix probeset (1570561_at, 22 probes including controls) actually represents the 16S-23S rRNA intergenic transcribed spacer of mycoplasma genomes. This DNA sequence was included in a human microarray (the HG-U133 +2) and so it measures expression of mycoplasma genes. Here we suggest many more individual HG-U133 +2 probes also do so and, *all* those that give a signal, are correlated. This correlation strengthens the earlier claim that a sample which express 1570561_at does so because it is contaminated by mycoplasma. Given the disruptive effect of mycoplasma on human cells' metabolisms [6], if a sample is contaminated no gene expression measurements from it (whether measured by microarray or any other technique) can be relied upon.

Whilst Aldecoa-Otalora *et al.* [8] suggested 1570561_at was the only *probeset* to target Mycoplasma arthritidis, firstly several more species of mycoplasma have been fully sequenced since, and secondly here we report many individual HG-U133 +2 probes map to one or more of the published mycoplasma genomes (see Table 2).

2 Method

As part of a study of alternative splicing of human exons previously we had down loaded, checked for spatial errors and quantile normalised all the human GeneChip CEL files from NCBI's GEO repository [9; 10; 11]. In particular we have 2757 HG-U133 +2, which are now available via RNAnet [12]. (GEO, like other Bioinformatics databases, continues to grow rapidly and it now contains many more data.)

Aldecoa-Otalora *et al.* [8] suggested that an EST DNA sequence within the reference human genome is actually DNA from mycoplasma. This public sequence was used by Affymetrix when they designed their HG-U133 +2 GeneChip and Aldecoa-Otalora *et al.* suggested that the probes in this probeset do not measure expression of human genes but instead they measure expression of mycoplasma genes. Using RNAnet they were able to show the probeset was essentially quiescent except in about 0.7% of GEO. Aldecoa-Otalora *et al.* suggested that in those cases the probeset was active because the supposedly human samples were in fact contaminated with mycoplasma. As support for this they also reported that the suspect samples were significantly more likely to be from cell lines [8, Supplementary Material].

Since [8] was published more species of mycoplasma have been fully sequenced. Using Bowtie [13] (release 0.12.7 with parameters `--all --best`) we find 437 HG-U133 +2 probes (including control probes) match one or more species of the 30 mycoplasma genomes we downloaded from `ftp.ncbi.nih.gov` (see the appendix). We restricted our search to the 106 probes that match one or more mycoplasma genome exactly. (None of these 106 are control probes.) We then calculated all possible pairwise correlations for individual probes. We report normal (i.e. Pearson) correlation but also calculated Spearman's rank correlation since it can be readily converted into a non-parametric statistical test. We also formulated a second null hypothesis by dividing at the median values each probe verses probe scatter plot into four quadrants. If there is no correlation between the two probes the four quadrants should contain approximately equal numbers of points. We test this with a Chi-squared test.

2.1 Setting a Signal Threshold

Microarray data are notoriously noisy. Although the RNAnet data have been filtered for spatial errors [11] some noise remains. Low intensity signals are especially prone to crosstalk from other nearby probes giving loud signals [14]. Figure 1 shows two example scatter plots where most of the data lie in the range 50–100 and are essentially noise. There are many reasons why a probe may give a weak signal, including poor

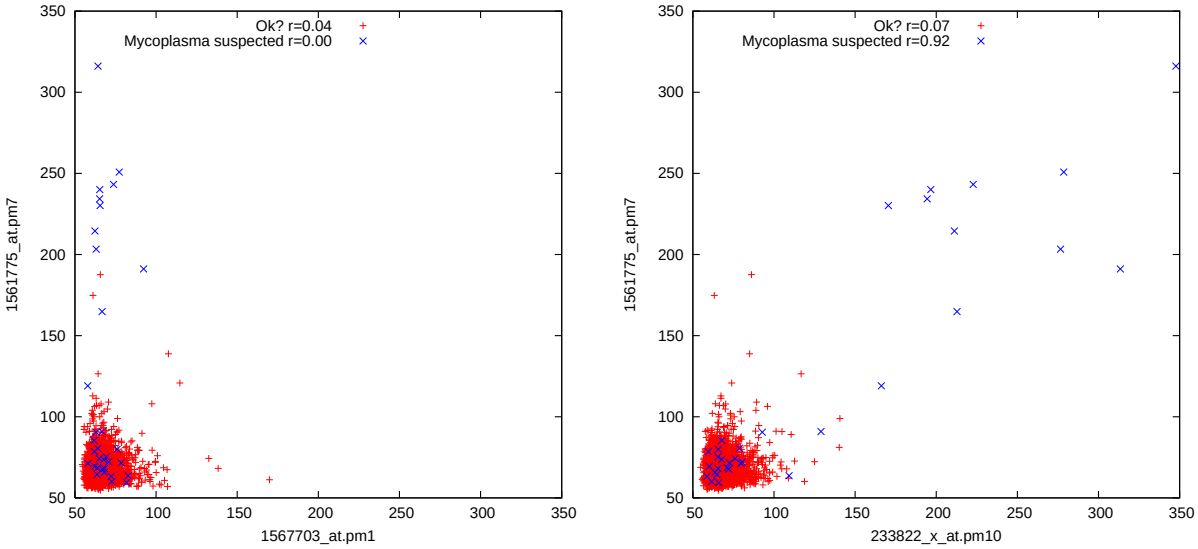

Figure 1: Sample scatter plots of normalised HG-U133 +2 probes which match published mycoplasma genome exactly. Gene expression values taken from GEO. 33 samples suspected of mycoplasma contamination plotted with $\times$. Left chosen as has small correlation. Right chosen with high correlation and same vertical data.

probe design. So below a certain signal strength we cannot rely on probe data. This section is concerned with setting a threshold below which we shall ignore probe signals.

We approached the problem of where to set a reasonable threshold on GeneChip data by noting that in general probes in the same probeset should be correlated. Therefore we took the subset of our all-pairs correlations corresponding to both pairs being in the same probeset and plotted them. In Figure 2 the horizontal axis is used simply to order all these correlations by the mean expression (of the least active of each pair).

It is clear from Figure 2 that when both probes are active (i.e. both have average expression above 120) then they are correlated. (53% of normalised probe values are below 120. See also Figure 3.) Therefor of our 106 probes which exactly match a mycoplasma genome we selected the 61 with mean normalised expression of at least 120. This gives $C_2^{61} = 1830$ pairings.

3 Results

Table 1 gives the correlations of all 61 probes which match exactly against mycoplasma and have a reasonable expression on the 33 suspect cel files. Apart from a small number of exceptions and the anomalous behaviour of one of the 61 probes (next paragraph) *all* are correlated.

Even at $p = 0.1$, only probe (211690_at.pm8) fails to show statistically significant correlation against many of the other 60 probes. We suggest 211690_at.pm8 is atypical because of two outliers. (The two outliers are shown in Figure 4. Apart from them 211690_at.pm8 does not have a large signal.) Hence we feel justified in excluding it. Except 211690_at.pm8, there are only 13 other pairs (shown in bold in Table 1) with poor statistical significance.

Like Table 1, the numbers in Figures 5 and 6 show Pearson correlation (multiplied by ten) for our 61 probes. White–Yellow backgrounds indicate high correlation, whilst blue indicates near zero. The high contrast between correlation on the contaminated data (Figure 5) and little correlation across several thousand other CEL files (Figure 6) is dramatic. (Figure 7 summarises Figures 5 and 6.)

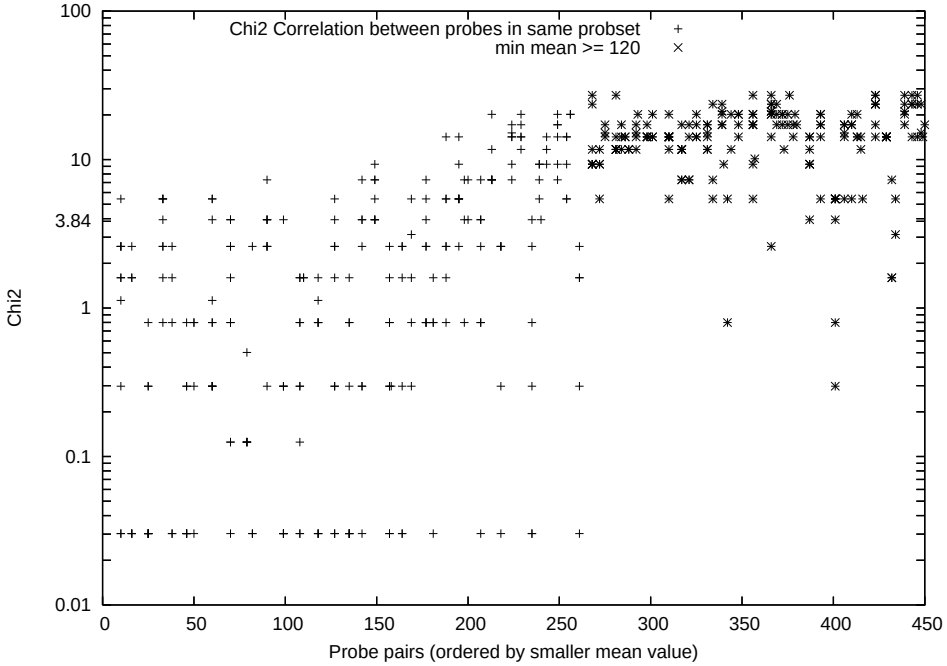

Figure 2: Using χ^2 to see which probe pairs in the same Affymetrix probeset are statistical correlated. Of the 106 HG-U133 +2 probes which match mycoplasma exactly, there are 450 pairs from the same probeset (plotted). Setting a threshold at 120 means only 7 pairs of 189 (3.7%) have $\chi^2 < 3.84$. ($\chi^2 > 3.84$ is needed for a p-value of better than 5%, 1 dof.)

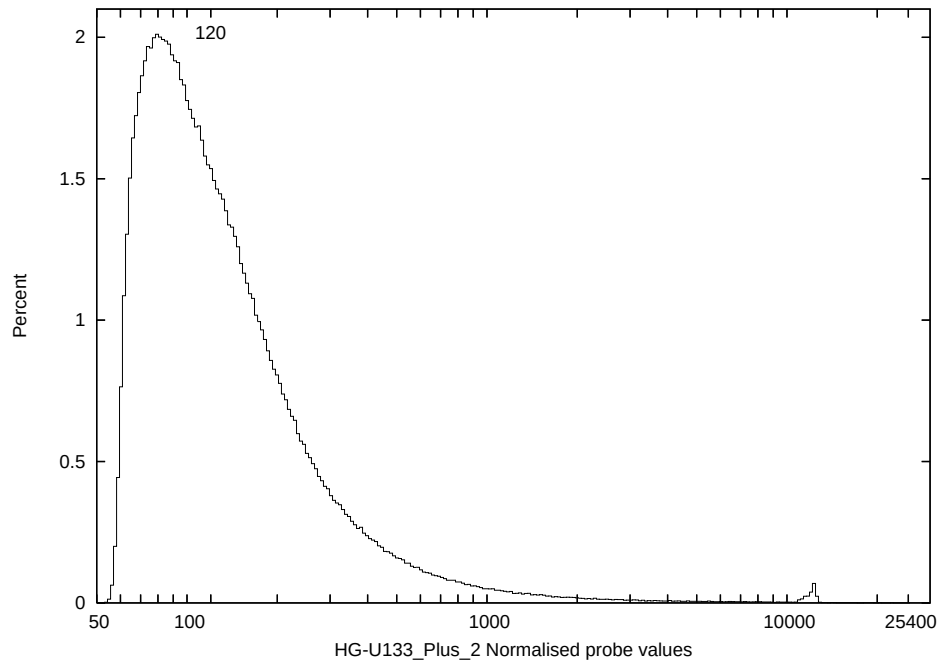

Figure 3: Distribution of normalised HG-U133 +2 probes. All GEO, centile bins, note log scale.

Table 1: Pearson correlation ($\times 10$) between HG-U133 +2 probes which match one or more species of mycoplasma genome exactly across 33 "Human" GEO cel files identified by [8] as suspicious. 3^{rd} column is location of the HG-U133 +2 probe in the Mycoplasma hyorhinis HUB-1 genome (NC_014448.1). 4^{th} column give average normalised expression in the 33 CEL files. Figures in bold indicate neither χ^2 more Spearman rank correlation show two probes to be statistically correlated at the 10% level. Excluding 211690_at.pm8 (see Figure 4), only 13 (of 1770, 0.7%) probe pairs fail to pass either statistical test.

	probe name	HUB-1	mean	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61
61	224354_at.pm1	752866	303	9	9	9	9	9	7	9	9	8	9	7	6	3	**1**	4	8	7	5	5	9	9	9	9	9	8	8	8	7	8	9	9	9	8	8	9	8	8	9	8	9	9	7	7	7	7	7	8	7	8	8	8	9	9	10	8	9	8	5	9	9	
60	224354_at.pm2	752849	380	8	9	9	9	10	6	9	9	9	9	6	5	2	**0**	2	7	8	4	5	9	8	9	8	9	8	8	9	8	8	9	8	9	8	8	8	8	7	8	8	8	8	7	7	7	7	7	8	8	8	8	8	10	10	9	9	10	8	5	9		9
59	224354_at.pm3	752835	176	8	9	9	9	9	6	9	9	8	8	6	7	4	**2**	2	8	7	5	6	8	9	9	9	9	8	8	9	7	8	8	8	9	7	8	8	8	8	8	8	8	8	8	8	7	7	8	8	8	9	9	8	9	9	10	8	9	7	6		9	9
58	224354_at.pm5	752768	490	5	5	5	6	5	6	6	5	4	4	3	8	8	**7**	1	8	6	8	7	3	3	3	2	3	2	1	3	2	2	3	3	4	**2**	3	2	2	2	2	3	3	3	1	2	4	5	8	7	6	7	7	7	4	5	5	4	5	3		6	5	5
57	224354_at.pm6	752748	248	6	8	8	8	8	**4**	7	7	7	7	9	4	0	-1	5	5	4	3	4	8	7	9	7	8	7	7	7	6	7	8	7	7	5	6	7	7	6	7	7	7	7	5	5	6	6	5	7	6	7	7	7	8	8	7	7	8		3	7	8	8
56	224354_at.pm7	752723	404	8	10	10	10	10	7	9	9	9	9	7	6	3	**1**	2	8	7	5	6	9	9	9	8	9	8	8	9	8	8	9	8	9	7	8	8	8	7	8	8	9	8	7	7	7	7	8	8	8	9	9	9	10	10	9	9		8	5	9	10	9
55	224354_at.pm8	752649	142	6	8	8	8	9	4	7	8	9	8	5	3	1	**0**	1	6	6	3	3	9	7	8	7	8	8	8	9	8	8	8	7	8	7	6	7	6	6	7	7	7	6	6	7	8	8	6	7	6	7	7	7	9	9	8		9	7	4	8	9	8
54	224354_at.pm9	752522	357	9	9	9	9	9	7	9	9	8	9	6	6	3	**1**	2	8	7	5	5	9	9	9	9	9	8	8	9	8	8	9	9	9	8	8	9	8	8	9	9	9	8	8	8	7	7	7	8	7	8	8	8	10	10		8	9	7	5	10	9	10
53	224354_at.pm10	752507	388	8	9	9	9	10	6	9	9	9	9	7	5	2	**1**	2	7	7	5	5	9	9	9	8	9	8	8	9	8	9	9	8	9	8	8	8	8	7	8	8	8	8	7	7	8	8	8	8	8	8	9	8	10		10	9	10	8	5	9	10	9
52	224354_at.pm11	752425	359	8	9	9	9	9	6	9	9	9	9	6	4	1	**-0**	2	7	7	3	4	10	9	9	9	9	9	9	9	9	9	9	9	9	8	8	9	8	8	9	9	9	8	8	8	7	7	6	7	7	7	8	7		10	10	9	10	8	4	9	10	9
51	1567703_at.pm4	661475	140	6	8	7	8	8	5	7	7	7	6	6	7	5	4	2	8	6	7	7	7	6	7	6	7	6	6	7	6	6	7	5	7	4	5	5	5	4	5	5	6	5	5	5	8	8	9	10	9	10	10		7	8	8	7	9	7	7	8	8	8
50	1567703_at.pm5	661460	191	6	8	7	8	8	5	7	7	7	6	6	7	5	**4**	2	8	7	7	6	7	6	7	6	7	6	6	7	6	6	7	6	7	4	5	5	5	5	5	6	6	6	6	6	7	7	9	10	10	10		10	8	9	8	7	9	7	7	9	8	8
49	1567703_at.pm6	661458	221	6	8	7	8	8	5	7	7	6	6	6	7	5	**4**	2	8	7	7	6	7	6	7	6	6	6	6	7	5	6	7	6	7	4	5	5	5	5	5	5	6	6	5	6	7	7	9	10	10		10	10	7	8	8	7	9	7	7	9	8	8
48	1567703_at.pm7	661453	250	5	7	7	8	7	4	7	6	6	6	6	6	5	**4**	1	8	7	7	6	6	5	7	5	6	5	6	6	5	6	6	5	6	4	4	5	4	4	4	5	5	5	5	5	7	7	9	10		10	10	9	7	8	7	6	8	6	6	8	8	7
47	1567703_at.pm8	661451	226	5	8	7	8	8	4	7	7	7	6	6	7	5	**4**	1	8	7	7	6	7	6	7	5	6	5	6	7	5	6	6	5	7	4	4	5	5	4	5	5	5	5	5	5	7	7	9		10	10	10	10	7	8	8	7	8	7	7	8	8	8
46	1567703_at.pm9	661448	157	5	7	6	7	7	4	7	6	6	5	4	8	7	**6**	1	8	6	6	6	6	5	6	5	5	4	5	6	5	5	5	4	6	4	4	4	4	4	4	5	5	4	5	5	7	7		9	9	9	9	9	6	8	7	6	8	5	8	8	7	7
45	1567703_at.pm10	661443	121	5	6	6	7	7	3	6	6	6	6	4	5	4	2	1	6	4	4	4	7	6	6	6	7	6	7	7	7	6	6	5	7	5	5	6	5	5	5	6	6	5	5	6	9		7	7	7	7	7	8	7	8	7	8	7	6	5	7	7	7
44	1567703_at.pm11	661440	123	4	6	6	6	7	3	6	6	7	6	5	4	3	1	1	6	4	4	4	7	5	6	5	6	7	7	7	7	7	6	5	6	4	4	5	4	4	4	5	5	4	5	5		9	7	7	7	7	7	8	7	8	7	8	7	6	4	7	7	7
43	233847_x_at.pm4	658617	215	5	5	6	6	6	2	6	6	6	6	4	3	0	**0**	1	5	6	1	0	8	8	7	9	8	8	8	9	8	9	8	7	8	8	8	8	7	8	8	8	7	8	10		5	6	5	5	5	6	6	5	8	7	8	7	7	5	2	8	7	7
42	233847_x_at.pm10	658441	308	6	6	7	6	6	3	7	7	6	7	4	3	0	**-0**	1	5	6	1	1	8	9	8	9	9	8	8	8	8	8	8	8	8	8	8	9	8	8	8	9	8	8		10	5	5	5	5	5	5	6	5	8	7	8	6	7	5	1	8	7	7
41	234623_x_at.pm2	458898	338	8	8	8	8	8	5	8	8	7	8	6	4	0	**-1**	3	5	6	3	3	9	10	9	9	9	8	8	8	8	8	9	10	9	9	10	10	9	10	10	10	10		8	8	4	5	4	5	5	6	6	5	8	8	8	6	8	7	3	8	8	9
40	234432_at.pm7	458777	128	9	8	9	8	9	6	9	9	8	9	6	4	0	-1	3	6	6	3	4	9	10	9	9	10	9	8	8	8	8	10	10	9	9	9	10	9	9	10	10		10	8	7	5	6	5	5	5	6	6	6	9	8	9	7	9	7	3	8	8	9
39	234432_at.pm6	458762	122	8	8	8	8	8	5	8	9	8	8	5	4	0	**-1**	2	6	6	2	3	9	10	9	9	9	9	8	9	8	8	9	9	10	9	10	10	9	10	10		10	10	9	8	5	6	5	5	5	5	6	5	9	8	9	7	8	7	3	8	8	8
38	234623_x_at.pm5	458703	175	9	8	8	8	8	6	8	9	8	9	5	4	-0	**-1**	2	5	6	2	3	9	10	9	9	9	8	7	8	8	8	9	10	9	10	10	10	10	10		10	10	10	8	8	4	5	4	5	4	5	5	5	9	8	9	7	8	7	2	8	8	9
37	234623_x_at.pm6	458667	162	8	7	8	7	8	6	8	8	7	8	4	4	0	**-0**	2	5	6	2	3	8	9	8	9	9	8	7	8	7	7	9	10	9	9	10	10	9		10	10	9	10	8	8	4	5	4	4	4	5	5	4	8	7	8	6	7	6	2	8	7	8
36	234623_x_at.pm7	458657	160	8	8	8	8	8	6	8	8	8	9	5	4	0	**-1**	2	5	5	2	3	9	9	9	9	9	7	7	8	7	8	9	10	9	9	9	10		9	10	9	9	9	8	7	4	5	4	5	4	5	5	5	8	8	8	6	8	7	2	8	8	8
35	234623_x_at.pm8	458639	204	8	8	8	8	8	5	8	9	8	9	5	4	0	**-1**	2	5	6	2	3	9	10	9	9	9	8	8	8	8	8	9	10	9	10	10		10	10	10	10	10	10	9	8	5	6	4	5	5	5	5	5	9	8	9	7	8	7	2	8	8	9
34	234623_x_at.pm9	458622	189	9	7	8	8	8	6	8	8	7	8	5	4	0	**-1**	2	5	6	2	3	8	10	8	9	9	8	7	8	7	7	9	10	9	9		10	9	10	10	10	9	10	8	8	4	5	4	4	4	5	5	5	8	8	8	6	8	6	3	8	8	8
33	234623_x_at.pm10	458613	149	7	7	8	7	8	5	7	8	8	8	4	3	-0	**-1**	1	5	6	1	**2**	9	9	8	9	9	8	7	8	8	8	9	9	9		9	10	9	9	10	9	9	9	8	8	4	5	4	4	4	4	4	4	8	8	8	7	7	5	**2**	7	8	8
32	234432_at.pm4	458586	127	8	8	9	8	9	5	8	9	9	9	6	4	0	**-0**	2	6	7	3	3	10	9	9	9	10	9	8	9	9	9	10	9		9	9	9	9	9	9	10	9	9	8	8	6	7	6	7	6	7	7	7	9	9	9	8	9	7	4	9	9	9
31	234623_x_at.pm11	458574	232	9	8	9	8	9	6	9	9	8	9	6	4	0	**-1**	2	6	6	3	4	9	10	9	9	9	8	8	8	8	8	9		9	9	10	10	10	10	10	9	10	10	8	7	5	5	4	5	5	6	6	5	9	8	9	7	8	7	3	8	8	9
30	234432_at.pm3	458574	169	8	9	9	9	9	5	9	9	9	9	7	4	0	**-1**	3	6	6	3	3	10	9	10	9	10	9	9	9	9	9		9	10	9	9	9	9	9	9	9	10	9	8	8	6	6	5	6	6	7	7	7	9	9	9	8	9	8	3	8	9	9
29	1561775_at.pm7	313206	121	6	7	8	7	8	**2**	7	7	8	7	5	2	-0	**-1**	2	5	6	1	**1**	9	8	8	9	9	9	9	10	9		9	8	9	8	7	8	8	7	8	8	8	8	8	9	7	6	5	6	6	6	6	6	9	9	8	8	8	7	2	8	8	8
28	1561775_at.pm6	313184	126	6	7	8	7	8	2	7	7	8	7	5	2	-0	-2	2	4	6	1	1	9	8	8	8	9	9	9	10		9	9	8	9	8	7	8	7	7	8	8	8	8	8	8	7	7	5	5	5	5	6	6	9	8	8	8	8	6	2	7	8	7
27	1561775_at.pm5	313181	164	6	7	8	7	8	3	7	8	8	8	6	4	0	-1	2	5	6	2	2	9	9	9	9	9	10	9		10	10	9	8	9	8	8	8	8	8	8	9	8	8	8	9	7	7	6	7	6	7	7	7	9	9	9	9	9	7	3	9	9	8
26	1561775_at.pm2	313114	162	6	7	7	7	8	2	7	7	7	7	6	2	-1	-2	2	4	6	1	1	9	8	9	9	9	9		9	9	9	9	8	8	7	7	8	7	7	7	8	8	8	8	8	7	7	5	6	6	6	6	6	9	8	8	8	8	7	1	8	8	8
25	1561775_at.pm1	313103	127	6	7	8	7	8	**3**	7	7	8	7	6	2	**-1**	-2	3	4	6	**0**	**1**	9	9	8	9	9		9	10	9	9	9	8	9	8	8	8	7	8	8	9	9	8	8	8	7	6	4	5	5	6	6	6	9	8	8	8	8	7	2	8	8	8
24	233822_x_at.pm4	185155	202	8	8	9	8	9	5	8	9	8	9	6	4	-0	-1	2	6	6	2	3	10	10	9	10		9	9	9	9	9	10	9	10	9	9	9	9	9	9	9	10	9	9	8	6	7	5	6	6	6	7	7	9	9	9	8	9	8	3	9	9	9
23	233822_x_at.pm7	185083	141	8	7	8	7	8	5	8	8	7	8	6	4	0	**-1**	3	5	6	2	2	9	10	9		10	9	9	9	8	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	5	6	5	5	5	6	6	6	9	8	9	7	8	7	2	9	8	9
22	233822_x_at.pm8	185070	567	8	9	9	9	9	6	9	9	9	9	7	4	0	**-0**	4	6	7	4	4	9	9		9	9	8	9	9	8	8	10	9	9	8	8	9	9	8	9	9	9	9	8	7	6	6	6	7	7	7	7	7	9	9	9	8	9	9	3	9	9	9
21	233822_x_at.pm9	185028	175	9	8	9	8	9	6	9	9	8	9	6	5	0	**-0**	3	6	7	3	3	9		9	10	10	9	8	9	8	8	9	10	9	9	10	10	9	9	10	10	10	10	9	8	5	6	5	6	5	6	6	6	9	9	9	7	9	7	3	9	8	9
20	233822_x_at.pm10	185012	128	7	8	9	9	9	5	8	9	9	9	6	4	0	**-1**	2	6	7	3	3		9	9	9	10	9	9	9	9	9	10	9	10	9	8	9	9	8	9	9	9	9	8	8	7	7	6	7	6	7	7	7	10	9	9	9	9	8	3	8	9	9
19	1570561_at.pm11	20275	468	6	7	6	7	6	8	6	6	5	5	3	7	6	**5**	**1**	7	4	9		3	3	4	2	3	**1**	1	2	1	**1**	3	4	3	**2**	3	3	3	3	3	3	4	3	1	0	4	4	6	6	6	6	6	7	4	5	5	3	6	4	7	6	5	5
18	1570561_at.pm10	20264	556	5	6	5	6	5	6	6	5	4	4	2	7	7	**7**	**0**	8	4		9	3	3	4	2	2	**0**	1	2	1	1	3	3	3	1	2	2	2	2	2	2	3	3	1	1	4	4	6	7	7	7	7	7	3	5	5	3	5	3	8	5	4	5
17	1570561_at.pm2	20137	7142	6	7	7	7	7	5	7	7	6	6	4	6	3	**3**	0	8		4	4	7	7	7	6	6	6	6	6	6	6	6	6	7	6	6	6	5	6	6	6	6	6	6	6	4	4	6	7	7	7	7	6	7	7	7	6	7	4	6	7	8	7
16	1570561_at.pm1	20126	3312	7	8	7	8	7	7	8	7	6	6	4	9	7	**6**	1		8	8	7	6	6	6	5	6	4	4	5	4	5	6	6	6	5	5	5	5	5	5	6	6	5	5	5	6	6	8	8	8	8	8	8	7	7	8	6	8	5	8	8	7	8
15	211690_at.pm9	19121	177	2	2	2	2	2	**1**	2	2	1	2	5	2	0	-0		1	0	**0**	**1**	2	3	4	3	2	3	2	2	2	2	3	2	2	1	2	2	2	2	2	2	3	3	1	1	1	1	1	1	1	2	2	2	2	2	2	1	2	5	1	2	2	4
14	211690_at.pm8	19095	168	**1**	**1**	**0**	**1**	**0**	**3**	**2**	**0**	**0**	**0**	-2	7	**9**		-0	**6**	**3**	**7**	**5**	**-1**	**-0**	**-0**	**-1**	-1	-2	-2	-1	-2	**-1**	**-1**	**-1**	**-0**	**-1**	**-1**	**-1**	**-1**	**-0**	**-1**	**-1**	-1	**-1**	**-0**	**0**	1	2	**6**	**4**	**4**	**4**	**4**	4	**-0**	**1**	**1**	**0**	**1**	-1	**7**	**2**	**0**	**1**
13	211690_at.pm7	19080	671	3	3	2	3	2	4	4	2	1	1	-0	8		**9**	0	7	3	7	6	0	0	0	0	-0	**-1**	-1	0	-0	-0	0	0	0	-0	0	0	0	0	-0	0	0	0	0	0	3	4	7	5	5	5	5	5	1	2	3	1	3	0	8	4	2	3
12	211690_at.pm5	19012	154	7	6	5	6	5	7	7	6	4	5	3		8	7	2	9	6	7	7	4	5	4	4	4	2	2	4	2	2	4	4	4	3	4	4	4	4	4	4	4	4	3	3	4	5	8	7	6	7	7	7	4	5	6	3	6	4	8	7	5	6
11	211690_at.pm1	18889	161	5	6	6	6	7	**4**	6	6	5	6		3	-0	-2	5	4	4	2	3	6	6	7	6	6	6	6	6	5	5	7	6	6	4	5	5	5	4	5	5	6	6	4	4	5	4	4	6	6	6	6	6	6	7	6	5	7	9	3	6	6	7
10	1555623_at.pm11	8754	442	9	10	10	9	10	8	9	10	9		6	5	1	**0**	2	6	6	4	5	9	9	9	8	9	7	7	8	7	7	9	9	9	8	8	9	9	8	9	8	9	8	7	6	6	6	5	6	6	6	6	6	9	9	9	8	9	7	4	8	9	9
9	1555623_at.pm10	8681	180	8	9	10	9	10	6	8	9		9	5	4	1	**0**	1	6	6	4	5	9	8	9	7	8	8	7	8	8	8	9	8	9	8	7	8	8	7	8	8	8	7	6	6	7	6	6	7	6	6	7	7	9	9	8	9	9	7	4	8	9	8
8	1555623_at.pm9	8666	1149	9	10	10	10	10	8	10		9	10	6	6	2	**0**	2	7	7	5	6	9	9	9	8	9	7	7	8	7	7	9	9	9	8	8	9	8	8	9	9	9	8	7	6	6	6	6	7	6	7	7	7	9	9	9	8	9	7	5	9	9	9
7	1555623_at.pm8	8651	1401	10	9	9	10	9	8		10	8	9	6	7	4	**2**	2	8	7	6	6	8	9	9	8	8	7	7	7	7	7	9	9	8	7	8	8	8	8	8	8	9	8	7	6	6	6	7	7	7	7	7	7	9	9	9	7	9	7	6	9	9	9
6	1555623_at.pm6	8466	318	9	8	7	8	7		8	8	6	8	**4**	7	4	**3**	**1**	7	5	6	8	5	6	6	5	5	**3**	2	3	2	**2**	5	6	5	5	6	5	6	6	6	5	6	5	3	2	3	3	4	4	4	5	5	5	6	6	7	4	7	**4**	6	6	6	7
5	1555623_at.pm5	8338	466	9	10	10	10		7	9	10	10	10	7	5	2	**0**	2	7	7	5	6	9	9	9	8	9	8	8	8	8	8	9	9	9	8	8	8	8	8	8	8	9	8	6	6	7	7	7	8	7	8	8	8	9	10	9	9	10	8	5	9	10	9
4	1555623_at.pm4	8321	918	9	10	10		10	8	10	10	9	9	6	6	3	**1**	2	8	7	6	7	9	8	9	7	8	7	7	7	7	7	9	8	8	7	8	8	8	7	8	8	8	8	6	6	6	7	7	8	8	8	8	8	9	9	9	8	10	8	6	9	9	9
3	1555623_at.pm3	8309	587	9	10		10	10	7	9	10	10	10	6	5	2	**0**	2	7	7	5	6	9	9	9	8	9	8	7	8	8	8	9	9	9	8	8	8	8	8	8	8	9	8	7	6	6	6	6	7	7	7	7	7	9	9	9	8	10	8	5	9	9	9
2	1555623_at.pm2	8290	1693	9		10	10	10	8	9	10	9	10	6	6	3	**1**	2	8	7	6	7	8	8	9	7	8	7	7	7	7	7	9	8	8	7	7	8	8	7	8	8	8	8	6	5	6	6	7	8	7	8	8	8	9	9	9	8	10	8	5	9	9	9
1	1555623_at.pm1	8276	1265		9	9	9	9	9	10	9	8	9	5	7	3	**1**	2	7	6	5	6	7	9	8	8	8	6	6	6	6	6	8	9	8	7	9	8	8	8	9	8	9	8	6	5	4	5	5	5	5	6	6	6	8	8	9	6	8	6	5	8	8	9

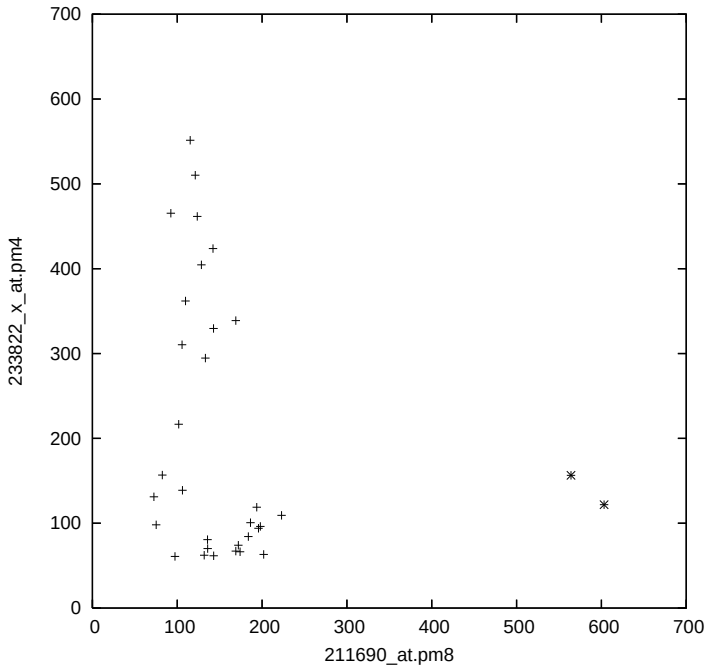

Figure 4: Scatter plot of normalised gene expression values for suspect GEO cel files for 211690_at.pm8 against another probe to show two outliers ($\star$ near $x = 600$).

4 Discussion

4.1 Does it Matter?

Contamination by mycoplasma is difficult to detect but the activity of mycoplasma genes can overwhelm the expressed RNA signal from human genes in the infected sample. Miller *et al.* [6] say mycoplasma contamination has "potentially major consequences for the diagnosis and characterization of diseases using expression array technology." Yet the suspect GEO data is used in five different publications in top flight journals. So far in total they have been cited 67 times. None of them explicitly mention mycoplasma contamination.

Only in one study are there a sizable number of published samples. In the others, it appears between 26% and 100% of the samples in the study were contaminated.

In all published cases the HG-U133 +2 measurements were backed up by real time PCR. Western blotting was also used in most cases. Although the publicly available data in GEO suggests the Affymetrix GeneChip samples were contaminated, other techniques are typically used to confirm HG-U133 +2 results and so are used later. This confirmation aims to overcome noise inherent in GeneChips and get more reliable measurements of expressed RNA rather than to address problems where the sample's metabolism has been changed by mycoplasma. Hence whilst we do not know that the samples used with RT-PCR etc. were also infected, there seems little reason to be confident that they were not.

4.2 Is it a Surprise?

Given the high frequency of mycoplasma contamination reported in microbiology laboratories (particularly for cell lines) [6], it is not unexpected that data from contaminated cell lines have been published. However, in addition to those previously reported, we find many Affymetrix probes designed from the human genome which match one or more published mycoplasma genomes and where they find a signal, they all respond in the same way giving, for GeneChips, unusually high correlations (see Figure 7).

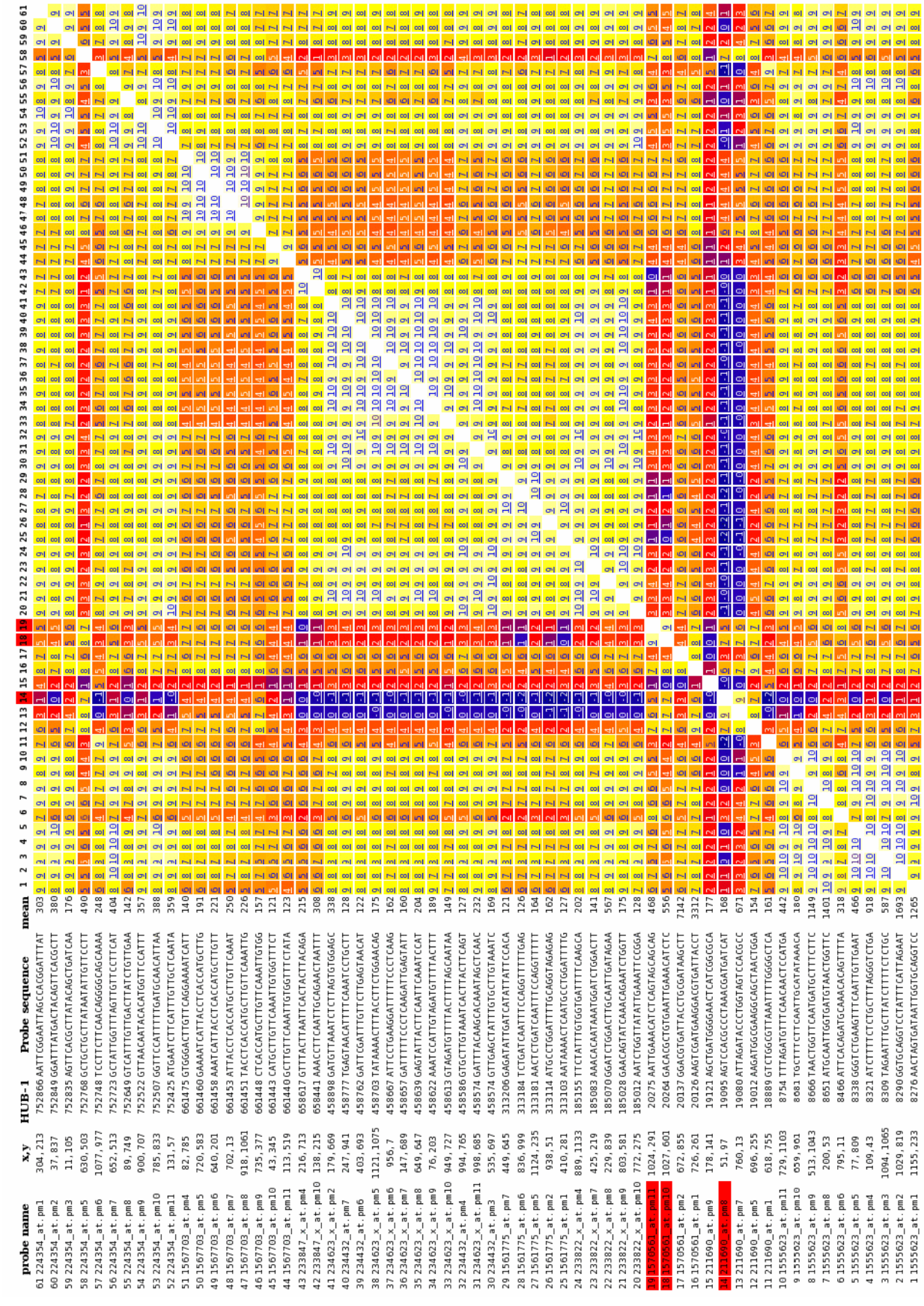

Figure 5: Pearson correlation (cf. Table 1) expressed as colours for 61 HG-U133 +2 Affymetrix probes which match mycoplasma exactly on 33 samples suspected of mycoplasma contamination

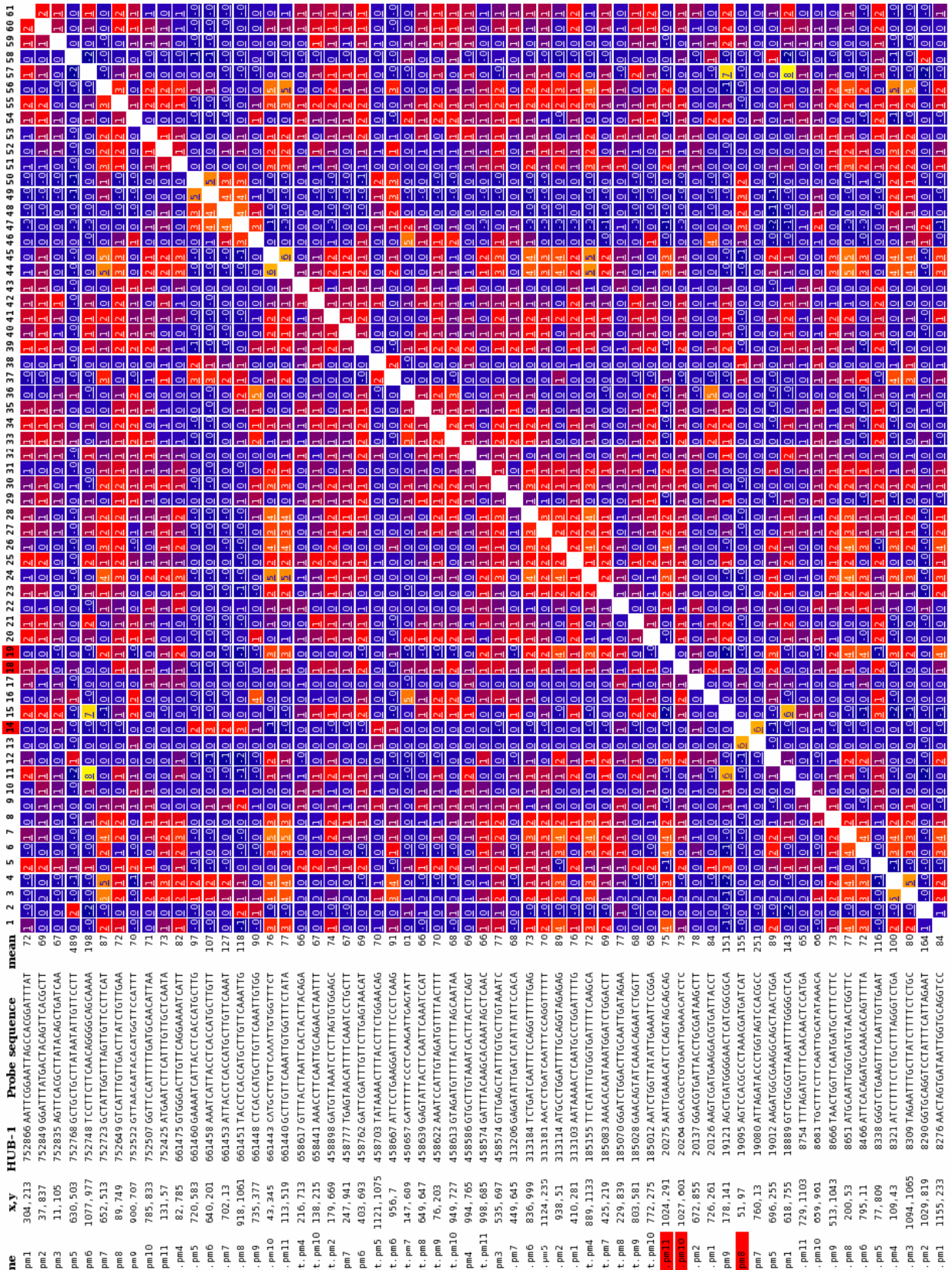

Figure 6: Pearson correlation expressed as colours (cf. Figure5) for same 61 HG-U133 +2 Affymetrix probes which match mycoplasma exactly across rest of HG-U133 +2 data in NCBI's GEO database. (Notice, as expected [15], probes with a run of 4 Gs are correlated.)

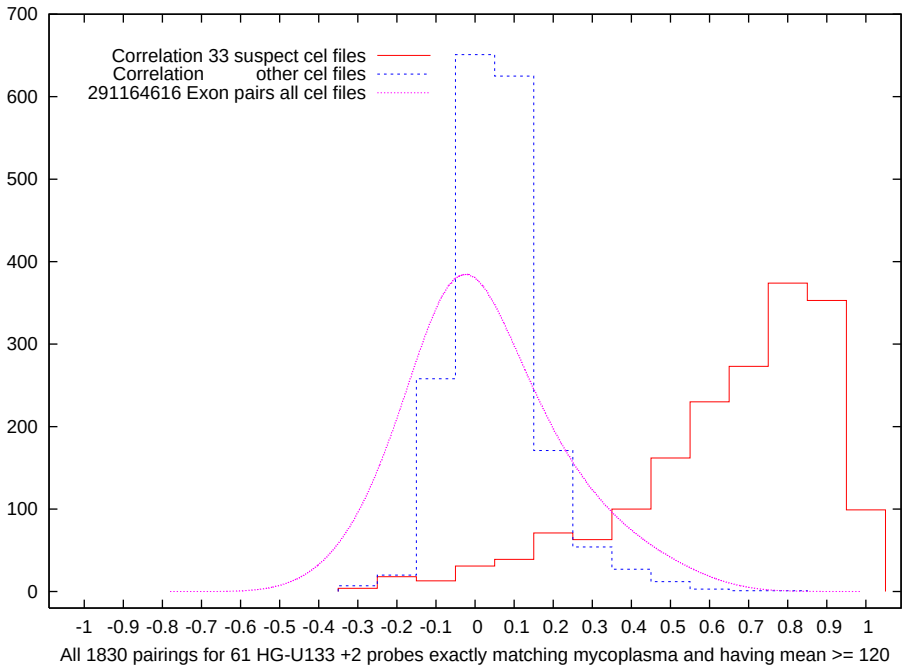

Figure 7: Distribution of all against all Pearson correlation coefficients for 61 HG-U133 +2 probes which map exactly to one or more species of mycoplasma bacteria and have a mean value ≥ 120. The two histograms contrast 33 human samples identified by [8] as suspected of being contaminated with mycoplasma (solid line, cf. Table 1 and Figure 5) v. rest of GEO (cf. Figure 6). (Data grouped in 0.1 wide bins). Dotted line is the underlying distribution. It is taken from 24132 well behaved HG-U133 +2 probes, one per Ensembl human exon (drawn to the same scale).

4.3 Correlation as an Investigative Bioinformatics Datamining Tool

The existence of NCBI's GEO and other large bioinformatics data repositories enables correlation studies like these which would be impractical for all but the largest laboratories or bioinformatics processing services. RNAnet provides convenient and near instant access to normalised GEO data and so allows cross site comparisons and data mining exploration of gene expression data. It has been used to investigate alternative exon splicing and alternative polyadenylation [16], human chimeric transcripts [17] and antisense expression (NAT) [18]. Given sufficient data, correlation is a powerful data mining tool. Other possibilities include cross correlating RNAnet (or other datasets) to investigate other contaminates, such as e-coli or viruses.

4.4 Errors in Public Datasets

Aldecoa-Otalora *et al.* [8] suggested a gene sequence in the human genome was not human but was in fact a DNA sequence from mycoplasma and further that it had been copied by a commercial company and incorporated into a gene expression measuring device (i.e. probeset 1570561_at on Affymetrix' HG-U133 +2 microarray). They also suggested that 33 public datasets in GEO are unreliable due to the presence of mycoplasma in the experiments they report. Since then we have reported a second mycoplasma gene sequence (DA466599) in the human genome [19; 20] and recently Longo *at al.* [21] reported other (non-human) public genome sequences appear to have have been contaminated with human genes. Here we have strengthened our claim that the 33 public datasets in GEO were contaminated by mycoplasma by reporting another 1530 data pairs which are correlated ($p = 5\%$) across the 33 suspect datasets.

Despite Aldecoa-Otalora *et al.* [8] having been published three years ago, Both the original sequence (AF241217) and the second one (DA466599 [20]) are still described as "Homo sapiens" within the NCBI reference human genome. In view of the exponential rise in genomic sequence data available via the Internet, everyone needs to be increasingly suspicious of public genomic databases.

References

[1] Carlos El Hader, Sandra Tremblay, Nicolas Solban, Denis Gingras, Richard Beliveau, Sergei N. Orlov, Pavel Hamet, and Johanne Tremblay, "HCaRG increases renal cell migration by a TGF-alpha autocrine loop mechanism," *Am J Physiol Renal Physiol*, vol. 289, no. 6, pp. F1273–F1280, Dec 2005.

[2] Stefan Schmidt, Johannes Rainer, Stefan Riml, Christian Ploner, Simone Jesacher, Clemens Achmller, Elisabeth Presul, Sergej Skvortsov, Roman Crazzolara, Michael Fiegl, Taneli Raivio, Olli A. Jnne, Stephan Geley, Bernhard Meister, and Reinhard Kofler, "Identification of glucocorticoid-response genes in children with acute lymphoblastic leukemia," *Blood*, vol. 107, no. 5, pp. 2061–2069, March 1 2006.

[3] Anatoly L. Mayburd, Alfredo Martlinez, Daniel Sackett, Huaitian Liu, Joanna Shih, Jordy Tauler, Ingalill Avis, and James L. Mulshine, "Ingenuity network-assisted transcription profiling: Identification of a new pharmacologic mechanism for MK886," *Clin Cancer Res*, vol. 12, no. 6, pp. 1820–1827, Mar 15 2006.

[4] Graham D. Jack, M. Carla Cabrera, Michael L. Manning, Stephen M. Slaughter, Malcolm Potts, and Richard F. Helm, "Activated stress response pathways within multicellular aggregates utilize an autocrine component," *Cellular Signalling*, vol. 19, no. 4, pp. 772–781, 2007.

[5] David Cappellen, Thomas Schlange, Matthieu Bauer, Francisca Maurer, and Nancy E. Hynes, "Novel c-MYC target genes mediate differential effects on cell proliferation and migration," *EMBO Rep*, vol. 8, no. 1, pp. 70–76, Jan 2007, European Molecular Biology Organization.

[6] Crispin J. Miller, Heba S. Kassem, Stuart D. Pepper, Yvonne Hey, Timothy H. Ward, and Geoffrey P. Margison, "Mycoplasma infection significantly alters microarray gene expression profiles," *BioTechniques*, vol. 35, no. 4, pp. 812–814, October 2003.

[7] Hans G. Drexler and Cord C. Uphoff, "Mycoplasma contamination of cell cultures: Incidence, sources, effects, detection, elimination, prevention," *Cytotechnology*, vol. 39, no. 2, pp. 75–90, 2002.

[8] Estibaliz Aldecoa-Otalora, William B. Langdon, Phil Cunningham, and Matthew J. Arno, "Unexpected presence of mycoplasma probes on human microarrays," *BioTechniques*, vol. 47, no. 6, pp. 1013–1016, December 2009.

[9] Andrew P. Harrison, Joanna Rowsell, Renata da Silva Camargo, William B. Langdon, Maria Stalteri, Graham J.G. Upton, and Jose M. Arteaga-Salas, "The use of Affymetrix GeneChips as a tool for studying alternative forms of RNA," *Biochemical Society Transactions*, vol. 36, pp. 511–513, 2008.

[10] Jose M. Arteaga-Salas, Harry Zuzan, William B. Langdon, Graham J. G. Upton, and Andrew P. Harrison, "An overview of image-processing methods for Affymetrix GeneChips," *Briefings in Bioinformatics*, vol. 9, no. 1, pp. 25–33, 2008.

[11] W. B. Langdon, G. J. G. Upton, R. da Silva Camargo, and A. P. Harrison, "A survey of spatial defects in Homo Sapiens Affymetrix GeneChips," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 7, no. 4, pp. 647–653, oct.-dec 2009.

[12] W. B. Langdon, Olivia Sanchez Graillet, and A. P. Harrison, "RNAnet a map of human gene expression," arXiv:1001.4263, 24 Jan 2010.

[13] Ben Langmead, Cole Trapnell, Mihai Pop, and Steven Salzberg, "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome," *Genome Biology*, vol. 10, no. 3, pp. R25, 2009.

[14] Graham J. G. Upton, Olivia Sanchez-Graillet, Joanna Rowsell, Jose M. Arteaga-Salas, Neil S. Graham, Maria A. Stalteri, Farhat N. Memon, Sean T. May, and Andrew P. Harrison, "On the causes of outliers in affymetrix genechip data," *Briefings in Functional Genomics & Proteomics*.

[15] William B. Langdon, Graham J. G. Upton, and Andrew P. Harrison, "Probes containing runs of guanine provide insights into the biophysics and bioinformatics of Affymetrix GeneChips," *Briefings in Bioinformatics*, vol. 10, no. 3, pp. 259–277, 2009.

[16] Olivia Sanchez-Graillet, Joanna Rowsell, William B. Langdon, Maria A. Stalteri, Jose M. Arteaga Salas, Graham J.G. Upton, and Andrew P. Harrison, "Widespread existence of uncorrelated probe intensities from within the same probeset on Affymetrix GeneChips," *Journal of Integrative Bioinformatics*, vol. 5, no. 2, pp. 98, 2008.

[17] Joanna Rowsell, Renata da Silva Camargo, William B. Langdon, Maria A. Stalteri, and Andrew P. Harrison, "Uncovering the expression patterns of chimeric transcripts using surveys of Affymetrix GeneChips," *Journal of Integrative Bioinformatics*, vol. 7, no. 3, pp. 137, 2010.

[18] Olivia Sanchez-Graillet, Maria A. Stalteri, Joanna Rowsell, Graham J.G. Upton, and Andrew P. Harrison, "Using surveys of affymetrix GeneChips to study antisense expression," *Journal of Integrative Bioinformatics*, vol. 7, no. 2, pp. 114, 2010.

[19] W. B. Langdon and M. J. Arno, "More mouldy data: Virtual infection of the human genome," Tech. Rep. RN/11/14, Department of Computer Science, University College London, London WC1E 6BT, UK, 14 June 2011.

[20] W. B. Langdon and M.J. Arno, "*In Silico* infection of the human genome," in *10th European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics, EvoBIO 2012*, Mario Giacobini, Leonardo Vanneschi, and William S. Bush, Eds., Malaga, Spain, 11-13 April 2012, vol. 7246 of *LNCS*, pp. 245–249, Springer Verlag.

[21] Mark S. Longo, Michael J. O'Neill, and Rachel J. O'Neill, "Abundant human DNA contamination identified in non-primate genome databases," *PLoS ONE*, vol. 6, no. 2, pp. e16410, 02 2011.

[22] Wei Liu, Liurong Fang, Sha Li, Qiang Li, Zhemin Zhou, Zhixin Feng, Rui Luo, Guoqing Shao, Lei Wang, Huanchun Chen, and Shaobo Xiao, "Complete genome sequence of mycoplasma hyorhinis strain HUB-1," *Journal of Bacteriology*, vol. 192, no. 21, pp. 5844–5845, Nov 2010.

A Mycoplasma Genomes Used

All the Mycoplasma genomes were down loaded from FTP site `ftp.ncbi.nih.gov` files `genomes/Bacteria/Mycoplasma_*` (30 files, 24 November 2011) and incorporated into a Bowtie EBWT database. See Table 2.

Table 2: Genomes of thirty species of Mycoplasma

Genome fasta description	Mycoplasma Complete Genome
gi|148377268|ref|NC_009497.1|	agalactiae PG2
gi|291319937|ref|NC_013948.1|	agalactiae chromosome
gi|193082772|ref|NC_011025.1|	arthritidis 158L3-1
gi|339320528|ref|NC_015725.1|	bovis Hubei-1 chromosome
gi|313678134|ref|NC_014760.1|	bovis PG45 chromosome
gi|83319253|ref|NC_007633.1|	capricolum subsp. capricolum ATCC 27343
gi|240047135|ref|NC_012806.1|	conjunctivae HRC/581 chromosome
gi|294155300|ref|NC_014014.1|	crocodyli MP145 chromosome
gi|308189587|ref|NC_014552.1|	fermentans JER chromosome
gi|319776738|ref|NC_014921.1|	fermentans M64 chromosome
gi|294660180|ref|NC_004829.2|	gallisepticum str. R(low) chromosome
gi|108885074|ref|NC_000908.2|	genitalium G37
gi|321309518|ref|NC_014970.1|	haemofelis str. Langford 1
gi|269114774|ref|NC_013511.1|	hominis ATCC 23114 chromosome
gi|54019969|ref|NC_006360.1|	hyopneumoniae 232
gi|72080342|ref|NC_007332.1|	hyopneumoniae 7448 chromosome
gi|71893359|ref|NC_007295.1|	hyopneumoniae J chromosome
gi|304372805|ref|NC_014448.1|	hyorhinis HUB-1 chromosome
gi|313664890|ref|NC_014751.1|	leachii PG50 chromosome
gi|47458835|ref|NC_006908.1|	mobile 163K
gi|330370665|ref|NC_015407.1|	mycoides subsp. capri LC str. 95010 plasmid pMmc-95010, complete sequence
gi|331703020|ref|NC_015431.1|	mycoides subsp. capri LC str. 95010
gi|127763381|ref|NC_005364.2|	mycoides subsp. mycoides SC str. PG1 chromosome
gi|26553452|ref|NC_004432.1|	penetrans HF-2
gi|13507739|ref|NC_000912.1|	pneumoniae M129
gi|15828471|ref|NC_002771.1|	pulmonis UAB CTIP
gi|344204770|ref|NC_015946.1|	putrefaciens KS1 chromosome
gi|325972867|ref|NC_015155.1|	suis str. Illinois chromosome
gi|325989358|ref|NC_015153.1|	suis KI3806
gi|71894025|ref|NC_007294.1|	synoviae 53

B HG-U133 +2 Probesets working with Mycoplasma

Table 3: Of the 473 HG-U133 +2 probes which match mycoplasma, 106 probes match one or more mycoplasma genomes (See Table 2) exactly. Of these 61 have a strong signal and come from the following ten probesets. (GO annotations are from Affymetrix' netaffy and so assume human genes.) Mycoplasma hyorhinis HUB-1 [22] gene ids from NCBI NC_014448.1

Probeset	GO biological process term	GO molecular function term	Symbol	HUB-1 Gene description
224354_at	glucose metabolic process, oxidation reduction	glyceraldehyde-3-phosphate dehydrogenase (phosphorylating) activity protein binding NAD or NADH binding	gap	Glyceraldehyde 3-phosphate dehydrogenase C
1567703_at			rpmF	50S ribosomal protein L32
233847_x_at			ribF	Riboflavin biosynthesis protein
234623_x_at			*as 234432_at*	
234432_at			MHR_0358	hypothetical protein
1561775_at			MHR_0246	hypothetical protein
233822_x_at	tRNA aminoacylation for protein translation	nucleotide binding aminoacyl-tRNA ligase activity ATP binding	serS	Seryl-trna synthetase protein
1570561_at	*first reported mycoplasma probeset [8]*		16S-23S ribosomal RNA intergenic spacer. I.e. lies between MHR_r001 16S ribosomal RNA (18326..19674) and MHR_r002 23S ribosomal RNA (20595..22932)	
211690_at	rRNA processing translation translational elongation TOR signaling cascade ribosomal small subunit biogenesis glucose homeostasis positive regulation of apoptosis	structural constituent of ribosome protein binding	MHR_r001	16S ribosomal RNA
1555623_at	oxidation reduction	oxidoreductase activity FAD or FADH2 binding	MHR_0008	dihydrolipoamide dehydrogenase