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

W. B. Langdon
Dept. of Computer Science, University College London Gower Street, London WC1E 6BT, UK
W.Langdon@cs.ucl.ac.uk

ABSTRACT

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 in homo sapiens studies between different microbiology laboratories in different countries which suggests contamination with mycoplasma is the common factor. This high lights the problem of experts in evolutionary computation needing to apply due diligence before relying on public medical datasets. *Caveat emptor* even if the data are free!

Categories and Subject Descriptors

I.2.8 [**Artificial Intelligence**]: [Problem Solving, Control Methods, and Search-heuristic methods]; J.3 [**Life and Medical Sciences**]: [Biology and genetics]; H.2.8 [**Database Applications**]: [Scientific databases]

General Terms

Data mining, Correlation and regression analysis

Keywords

DNA gene expression, Homo sapiens genome reference consortium GRCh37.p5 h_sapiens_37.5_asm, evolutionary algorithm

1. INTRODUCTION

Some mycoplasma species cause human diseases, such as a variety of respiratory syndromes and septic abortion and amnionitis [6], but they are not susceptible to penicillin [7]. Mycoplasma are the smallest bacteria, making them hard to see with a microscope. Generally they have small genomes, which are readily sequenced. This was one of the reasons why Craig Venter choose mycoplasma genitalium's genome (only half a million DNA bases) as the start for the first

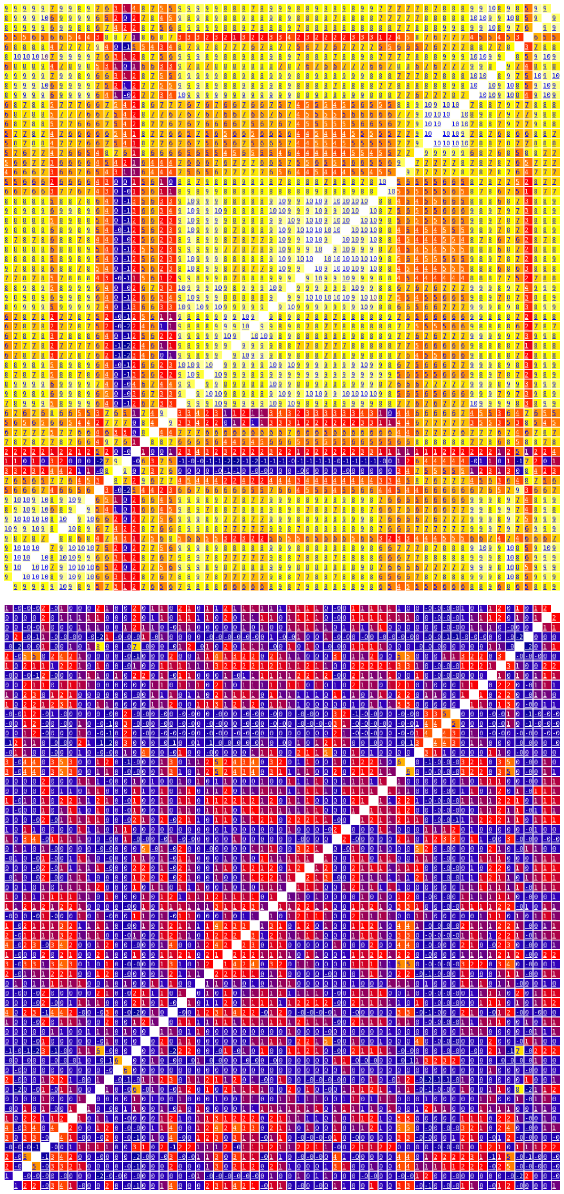

Figure 1: Correlation 61 human genome U133 +2.0 GeneChip probes which match mycoplasma exactly. Top: 33 samples suspected of mycoplasma contamination. Bottom: 2724 remaining HG-U133 +2 data in NCBI's GEO database. (Yellow high correlation.)

[*] Extends UCL technical report RN/12/11

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

GECCO'13 Companion, July 6–10, 2013, Amsterdam, The Netherlands.
Copyright 2013 ACM 978-1-4503-1964-5/13/07 ...$10.00.

artificial organism [8] (known as mycoplasma laboratorium). Not only are they ubiquitous in the human genitourinary tract [6] but readily grow in cultures used in microbiology laboratories [9]. Indeed their genes have recently crossed the silicon barrier and infect major Bioinformatics databases and tools [10, 11]. We will concentrate upon a new tool for analysing the consequent corruption of wet labs published data and potential challenges for evolutionary computation.

It is well known that mycoplasma contamination is rife in molecular biology laboratories [12, 13]. Depending upon medium, mycoplasma contamination rates of 1% to 15–35% (or even higher) have been reported [9]. Many labs routinely sterilised their equipment to counter it. About 1% of published NCBI's GEO GeneChip data appear to be contaminated [10]. Aldecoa-Otalora *et al.* [10] 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. Indeed mycoplasma contamination has been confirmed in some cases [13, 14]. Given the disruptive effect of mycoplasma on human cells' metabolisms [12], if a sample is contaminated no gene expression measurements from it (whether measured by microarray or any other technique) can be relied upon.

Although Aldecoa-Otalora *et al.* [10] suggested 1570561_at was the only *probeset* to target Mycoplasma arthritidis, this is not the full story. Here we report many HG-U133 +2 probes (i.e. members of probesets) map to one or more of the published mycoplasma genomes (see [15, Table 2]). (Also more species of mycoplasma have been fully sequenced.)

The next section discusses challenges for evolutionary algorithms when using diverse public gene expression datasets (e.g. RNAnet). This is followed (in Sections 3 to 5) by detailed descriptions of correlation across hundreds of published gene expression experiments. Finally Section 6 discusses the implications of our findings and the risks of naive use of medical data.

2. CHALLENGES FOR GENETIC AND EVOLUTIONARY COMPUTATION IN COMPUTATIONAL GENOMICS

Historically Bioinformatics has tended to view DNA only as a precursor to proteins. However with the exponential growth in published DNA sequences there is increasing realisation that DNA sequences, other than that coding for genes and hence proteins, plays a variety of important roles.

Evolutionary computing has already been used in computational genomics [16]. Burgeoning areas include: classifying medical tissues [17], medical imaging (which is particularly susceptible to parallel processing on GPUs [18]), the study of gene-gene interaction and gene regulatory networks (GRNs) [19], and recognising DNA sequences for gene promoters [20]. Evolutionary computation has also been used in data mining [21, 22] single nucleotide polymorphisms (SNPs) [23], epistasis [24] and phylogenetic studies [25, 26, 27]. EC phylogenetic studies are pleasingly circular. In that, a computational technique inspired by natural evolution (EC), is used to create evolutionary trees, which describe the creation of new species and species extinctions. EC uses the commonalities and the differences between current (and sometimes recent) genomes to infer phylogenetic trees which describe Life's ancestry.

Datamining of large public databases of gene expression levels, such as EBI's ArrayExpress and NCBI's GEO, opens the way to whole genome studies of gene-gene interaction. Previously gene-gene studies were driven by the need to test hypothesises created by experts. These allowed detailed studies of only a few genes. Experimental whole genome studies are beyond the resources of any single wet lab. Instead datamining of global datasets using modern search heuristics potentially allows the automatic data-driven generation of hypothesis. However these data are now remote from the original experiment and so interesting hypothesises will require expert confirmation and perhaps further traditional microbiology experiments. RNAnet [28] provides immediate on-line correlation between *all* human genes across many tissues and disease states. (Previously, even where feasible, just normalising these data from so many different publications took many hours.) Indeed we have precalculated correlation between 29637 representative gene expression data covering most human genes. (Since correlation is symmetric, we need only store $n(n-1)/2$ data, where n is the number of genes.) At present it appears that tools like GeNet [19] experience practical difficulties in dealing with 439 161 066 pair-wise correlations. However RNAnet can calculate subsets of these. With improved techniques and larger computers datamining all the human genes should be feasible. Indeed, like DNA sequences, the volume of public gene expression data continues to grow exponentially, demanding efficient techniques, like evolutionary algorithms to extract information from these rich data pools. Future work may consider three-way or even higher gene interactions.

Similar public data are available for other species. While these too are set to grow exponentially, they will probably always lag behind those for man. In many cases tools (e.g. GAs) developed for human data are suitable for many other organisms. Some, particularly plants, have more genes than man but even so homo sapiens tools and data analysis techniques can probably be readily scaled to many of them.

3. 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 [29, 30, 31]. In particular we have 2757 HG-U133 +2, which are now available via RNAnet [28]. (GEO, like other Bioinformatics databases, continues to grow rapidly and it now contains many more data.)

In Aldecoa-Otalora *et al.* [10] we suggested that an expressed sequence tag (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 we were able to show the probeset was essentially quiescent except in about 0.7% of GEO. Aldecoa-Otalora *et al.* suggested

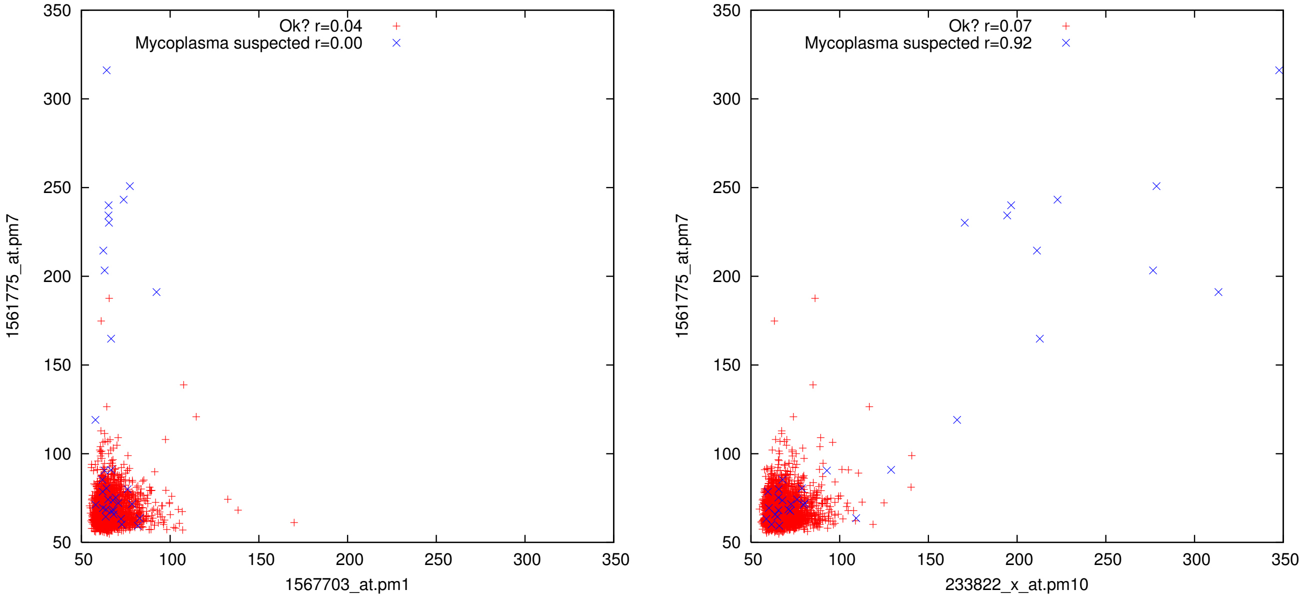

Figure 2: Sample scatter plots of normalised HG-U133 +2 probes which match published mycoplasma genome exactly. Gene expression values taken from GEO. 33 samples suspected of mycoplasma contamination plotted with ×. Left chosen as has small correlation. Right chosen with high correlation and same vertical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

that in those cases the probeset was active because the supposedly human samples were in fact contaminated with mycoplasma. As support for this we also reported that the suspect samples were significantly more likely to be from cell lines [10, Supplementary Material].

Since [10] was published more species of mycoplasma have been fully sequenced. Using Bowtie [32] (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 [15, 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 χ^2 Chi-squared test.

3.1 Setting the Signal Threshold at 120

Microarray data are notoriously noisy. Although the RNAnet data have been filtered for spatial errors [31] some noise remains. Low intensity signals are especially prone to crosstalk from other nearby probes giving loud signals [33]. Figure 2 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 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.

Probes in the same probeset should be correlated. Although the probesets were designed w.r.t. the reference human genome, the grouping of these probes into probesets also holds for mycoplasma. I.e. when probes from a "human" probeset are mapped against the mycoplasma hyorhinis HUB-1 [34] they still lie close to each other (and in the same order) and individual probesets map to individual HUB-1 genes. Therefore we took the subset of our all-pairs correlations corresponding to both pairs being in the same probeset and plotted them. In Figure 3 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 3 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 4.) 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.

4. 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 5. 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.

Figure 1 shows the Pearson correlation between our 61 probes. White–yellow backgrounds indicate high correlation, whilst blue indicates near zero. (Code to convert corre-

Table 1: Pearson correlation (×10) between HG-U133 +2 probes which match one or more species of mycoplasma genome exactly across 33 "Human" GEO cel files identified by [10] as suspicious. 3rd column is location of the HG-U133 +2 probe in the Mycoplasma hyorhinis HUB-1 genome (NC_014448.1). 4th column give average normalised expression in the 33 CEL files. Figures in bold indicate neither χ^2 nor Spearman rank correlation show two probes to be statistically correlated at the 10% level. Excluding 211690_at.pm8 (see Figure 5), 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

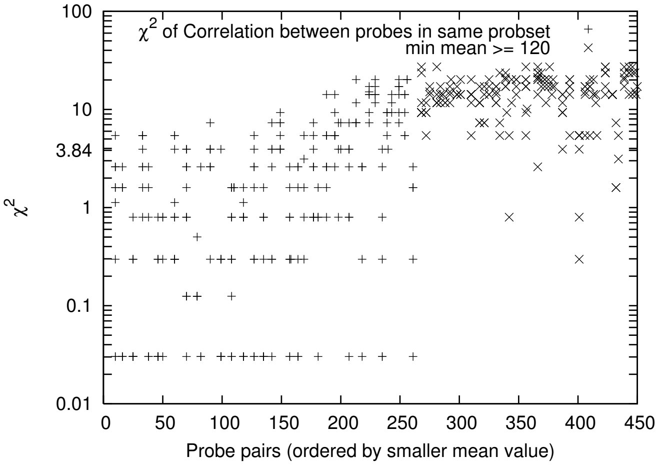

Figure 3: Using χ^2 to see which probe pairs in the same Affymetrix probeset are statistically correlated. Of the 106 HG-U133 +2 probes which match mycoplasma exactly, there are 450 pairs from the same probeset (plotted + and ×). Setting a threshold above 120 (×) gives only seven pairs (of 189, 3.7%) which have χ^2 below 3.84. (χ^2 above 3.84 is needed for a p-value of better than 5%, 1 dof.)

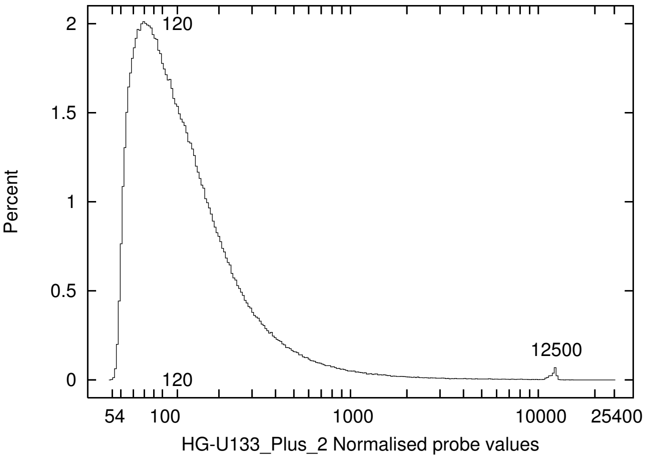

Figure 4: Distribution of normalised HG-U133 +2 probes. All GEO, centile bins, note log scale.

lation to a colour can be found via `http://bioinformatics.essex.ac.uk/users/wlangdon/colour.html`) The high contrast between correlation on the contaminated data (Figure 1 top) and little correlation across several thousand other CEL files (bottom of Figure 1) is dramatic. (Figure 6 summarises Figure 1.)

Figure 1 contains a few examples of badly behaved probes. These contain runs of four or more Gs. As expected [35], these are indeed correlated with other probes containing runs of Gs but not with other members of their probesets.

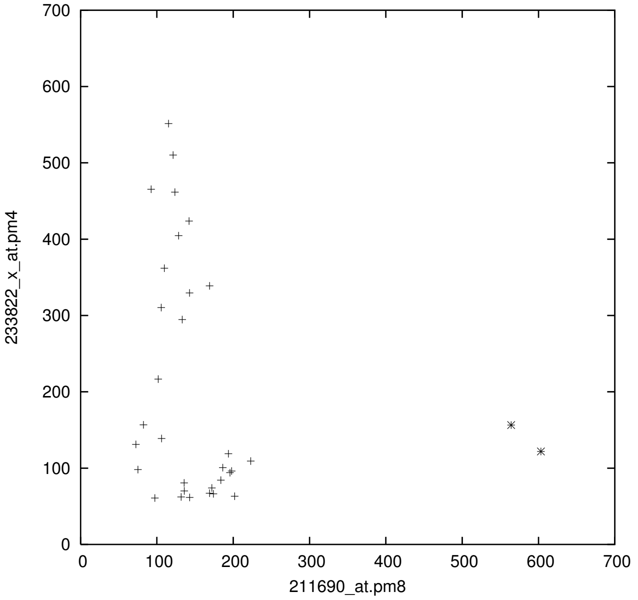

Figure 5: Scatter plot of normalised gene expression values for suspect GEO cel files for 211690_at.pm8 against another probe to show two outliers ($\star$ near $x = 600$).

5. INSUFFICIENT HG-U133A DATA

Although it is known that some of GEO's samples which correlation picked out as being suspect of mycoplasma contamination were indeed contaminated [13], we also sought independent confirmation from Miller *et al.* [12]. Unfortunately there were two problems: 1) Miller's group used an older microarray, Affymetrix's HG-U133A. 2) The data set includes only three GeneChips which were contaminated with mycoplasma.

Only one probeset in HG-U133A contains probes which match against one or more species of Mycoplasma exactly. It is probeset 211690_at. Miller *et al.* report that 211690_at is neither up nor down regulated by mycoplasma contamination. Also only only three of the eleven 211690_at probes (pm5, pm7 and pm8) are in our set of 60 Mycoplasma indicator probes.

We downloaded all Miller *et al.* [12]'s 12 CEL files, checked them for spatial errors and quantile normalised [31] [1] them using RNAnet's [28] robust average of HG-U133A gene expression from GEO. The signal for probeset 211690_at is weak and perhaps consequentially, even across all 11 PM probes the correlation is low (on average only 0.058694). So we found for Miller *et al.* [12]'s data 211690_at probes pm5 pm7 and pm8 did not correlate well with Mycoplasma presence/absence.

Naturally this is disappointing but given that we are looking at only three probes, on three contaminated samples, with low signal strength, perhaps not too surprising.

[1] `ftp://ftp.cs.ucl.ac.uk/genetic/gp-code/R`

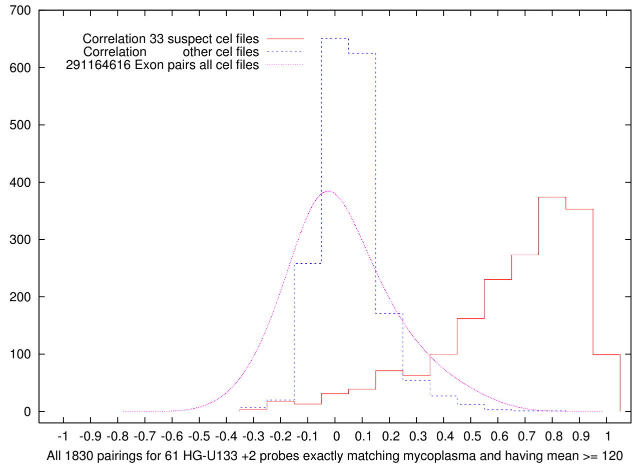

Figure 6: Distribution of all against all Pearson correlation coefficients for 61 HG-U133 +2 probes which map exactly to one or more species of mycoplasma bacteria and have a mean value ≥ 120. The two histograms contrast 33 human samples identified by [10] as suspected of being contaminated with mycoplasma (solid line, cf. Table 1 and top of Figure 1) v. rest of GEO (cf. bottom of Figure 1). (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).

6. DISCUSSION

6.1 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.* [12] 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. According to PubMed, so far in total they have been cited 68 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, except in one case [13], 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.

To some extent the importance of mycoplasma contamination has been accepted and this can be reflected in improved laboratory rigour [13].

6.2 Is it a Surprise?

Given the high frequency of mycoplasma contamination reported in microbiology laboratories (particularly for cell lines) [12], 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 6).

6.3 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 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 [36], human chimeric transcripts [37] and antisense expression (NAT) [38]. 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.

6.4 Errors in Public Datasets

In Aldecoa-Otalora *et al.* [10] we suggested a gene sequence in the human genome was not human but was in fact a DNA sequence from mycoplasma. 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.) We 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 [39, 11] and recently Longo *at al.* [40] 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.* [10] having been published three years ago, Both the original sequence (AF241217) and the second one (DA466599 [11]) 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.

Acknowledgements

I would like to thank Johannes Rainer, David Cappellen and Reinhard Kofler for helpful comments on technical report RN/12/11 [15] and Kevin Bryson for helpful discussions.

7. 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 Achmueller, Elisabeth Presul, Sergej Skvortsov, Roman Crazzolara, Michael Fiegl, Taneli Raivio, Olli A. Jaenne, 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] Herbert J. Harwick, George M. Kalmanson, and Lucien B. Guze, “Human diseases associated with mycoplasmas,” *California Medicine*, vol. 116, no. 5, pp. 1–7, May 1972.

[7] David Taylor-Robinson and Christiane Bebear, “Antibiotic susceptibilities of mycoplasmas and treatment of mycoplasmal infections.,” *Journal of Antimicrobial Chemotherapy*, vol. 40, no. 5, pp. 622–630, 1997.

[8] Daniel G. Gibson, Gwynedd A. Benders, Cynthia Andrews-Pfannkoch, Evgeniya A. Denisova, Holly Baden-Tillson, Jayshree Zaveri, Timothy B. Stockwell, Anushka Brownley, David W. Thomas, Mikkel A. Algire, Chuck Merryman, Lei Young, Vladimir N. Noskov, John I. Glass, J. Craig Venter, Clyde A. Hutchison, and Hamilton O. Smith, “Complete chemical synthesis, assembly, and cloning of a mycoplasma genitalium genome,” *Science*, vol. 319, no. 5867, pp. 1215–1220, 2008.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] David Cappellen, personal communication, 30 Nov 2012.

[14] Reinhard Kofler, personal communication, 15 May 2013.

[15] W. B. Langdon, “Correlation of microarray probes give evidence for mycoplasma contamination in human studies,” Tech. Rep. RN/12/11, Department of Computer Science, University College London, London WC1E 6BT, UK, 2 November 2012.

[16] Mohammad Wahab Khan and Mansaf Alam, “A survey of application: Genomics and genetic programming, a new frontier,” *Genomics*, vol. 100, no. 2, pp. 65–71, Aug. 2012.

[17] Michael A. Lones, Stephen L. Smith, Andrew T. Harris, Alec S. High, Sheila E. Fisher, D. Alastair Smith, and Jennifer Kirkham, “Discriminating normal and cancerous thyroid cell lines using implicit context representation cartesian genetic programming,” in *2010 IEEE World Congress on Computational Intelligence*, Pilar Sobrevilla, Ed., Barcelona, 18-23 July 2010, pp. 1945–1950, IEEE.

[18] Leonardo Vanneschi, Luca Mussi, and Stefano Cagnoni, “Hot topics in evolutionary computation,” *Intelligenza Artificiale*, vol. 5, no. 1, pp. 5–17, 2011.

[19] Leonardo Vanneschi, Matteo Mondini, Martino Bertoni, Alberto Ronchi, and Mattia Stefano, “GeNet: A graph-based genetic programming framework for the reverse engineering of gene regulatory networks,” 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 Apr. 2012, vol. 7246 of *LNCS*, pp. 97–109, Springer Verlag.

[20] Daniel Howard and Karl Benson, “Evolutionary computation method for promoter site prediction in DNA,” in *Genetic and Evolutionary Computation – GECCO-2003*, E. Cantú-Paz, J. A. Foster, K. Deb, D. Davis, R. Roy, U.-M. O’Reilly, H.-G. Beyer, R. Standish, G. Kendall, S. Wilson, M. Harman, J. Wegener, D. Dasgupta, M. A. Potter, A. C. Schultz, K. Dowsland, N. Jonoska, and J. Miller, Eds., Chicago, 12-16 July 2003, vol. 2724 of *LNCS*, pp. 1690–1701, Springer-Verlag.

[21] Stephan M. Winkler, Michael Affenzeller, and Stefan Wagner, "Using enhanced genetic programming techniques for evolving classifiers in the context of medical diagnosis," *Genetic Programming and Evolvable Machines*, vol. 10, no. 2, pp. 111–140, June 2009.

[22] W. B. Langdon and A. P. Harrison, "GP on SPMD parallel graphics hardware for mega bioinformatics data mining," *Soft Computing*, vol. 12, no. 12, pp. 1169–1183, Oct. 2008, Special Issue on Distributed Bioinspired Algorithms.

[23] Jason H. Moore and Bill C. White, "Exploiting expert knowledge in genetic programming for genome-wide genetic analysis," in *Parallel Problem Solving from Nature - PPSN IX*, Thomas Philip Runarsson, Hans-Georg Beyer, Edmund Burke, Juan J. Merelo-Guervos, L. Darrell Whitley, and Xin Yao, Eds., Reykjavik, Iceland, 9-13 Sept. 2006, vol. 4193 of *LNCS*, pp. 969–977, Springer-Verlag.

[24] Jason H. Moore, Nate Barney, Chia-Ti Tsai, Fu-Tien Chiang, Jiang Gui, and Bill C. White, "Symbolic modeling of epistasis," *Human Heredity*, vol. 63, no. 2, pp. 120–133, Feb. 2007.

[25] Clare Bates Congdon and Kevin J. Septor, "Phylogenetic trees using evolutionary search: Initial progress in extending gaphyl to work with genetic data," in *Proceedings of the 2003 Congress on Evolutionary Computation CEC2003*, Ruhul Sarker, Robert Reynolds, Hussein Abbass, Kay Chen Tan, Bob McKay, Daryl Essam, and Tom Gedeon, Eds., Canberra, 8-12 Dec. 2003, pp. 320–326, IEEE Press.

[26] Carlos Cotta and Pablo Moscato, "Inferring phylogenetic trees using evolutionary algorithms," in *Parallel Problem Solving from Nature - PPSN VII*, Juan J. Merelo-Guervos, Panagiotis Adamidis, Hans-Georg Beyer, Jose-Luis Fernandez-Villacanas, and Hans-Paul Schwefel, Eds., Granada, Spain, 7-11 Sept. 2002, number 2439 in Lecture Notes in Computer Science, LNCS, pp. 720–729, Springer-Verlag.

[27] Rudi Cilibrasi and Paul Vitanyi, "A new quartet tree heuristic for hierarchical clustering," in *Principled methods of trading exploration and exploitation Workshop*, London, 6-7 July 2005.

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

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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*, vol. 8, no. 3, pp. 199–212, 2009.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.