Advances in Medicine and Biology, 159 Leon V. Berhardt *Ed.*, Chapter 8, pp 225-238
Nova Science Publishers, Inc. 2020 *Preprint*

Mycoplasma ∪ Homo Sapiens: Contamination in E-Science

William B. Langdon

CREST, Computer Science, UCL, London, UK
W.Langdon@cs.ucl.ac.uk

Abstract. We are witnessing the big data explosion in Biology. Even before the publication of the "rough draft" of the human genome, the volume of sequence data was growing exponentially. For example, the leading Bioinformatics data repository, the USA's National Center for Biotechnology Information (NCBI) provides Internet access to petabytes (10^{15} bytes) of data. Free online access via the world wide web (WWW) to Bioinformatics data is essential to synthetic biology, *in silico* studies and has enabled rapid progress in both medicine and agriculture. However these computer based data are not perfect. We describe several cases where Mycoplasma have lead to electronic contamination of prestigious public data stores.

Keywords *in silico* experiments, E-contamination, data cleansing, sneaky bacteria, cell line, NCBI GEO, GenBank, 1KGP, next generation DNA sequencing (NGS), selfish silicon gene, HG-U133 +2 GeneChip

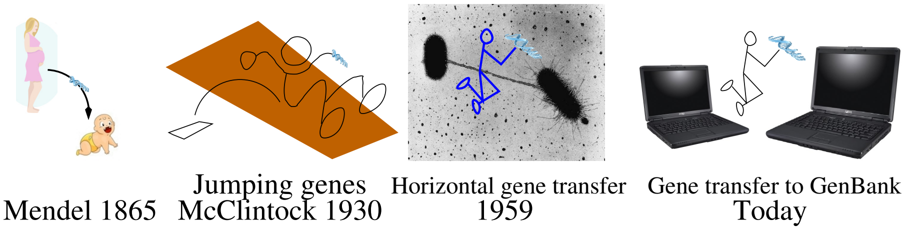

Fig. 1. Tetratych showing: 1865 Mendel's [1] discovery of the essential digital nature of inheritance. 1930 Barbara McClintock's [2] discovery of transposons in Maize whereby genes move not only from parent to child but also along chromosomes. 1959 Micrograph of genetic transfer along a pilus linking two bacteria. (Akiba and Ochia discovered the first interspecies gene transfer [3].) Mycoplasma bacteria genes are transferred between computers, including into the reference human genome DNA sequence held by GenBank [4].

1 Big Data need not be Correct

For convenience we take the start of the third millennium as the start of the big data explosion in Biology. But even before the year 2000 the volume of sequence data was growing exponentially and by 2015 the National Center for Biotechnology Information (NCBI) housed more than five petabytes of Bioinformatics data. Free access to such data enables rapid scientific and technological progress, however we will describe several cases where Mycoplasma and other species have lead to electronic contamination of prestigious public data stores.

Mycoplasma are the smallest free living organisms [5]. They are almost transparent and hence difficult to detect. Under ideal conditions, such as may be found in microbiology laboratories, they multiply and grow rapidly and their biological activity may totally overwhelm the gene expression signal from the supposed tissue sample [6,7]. It is now well know that they can contamination wet-ware experiments and many laboratories routinely sterilise all their glassware, even if no contamination has been detected. However this has not necessarily been universal practise in the past.

We will describe post-2000 documented cases where Mycoplasma has indeed contaminated human samples. This has meant that their DNA have been sampled alongside human genes and wrongly labelled as Homo sapiens. Examples include the reference human genome [4,8] and the first large scale human Single-nucleotide polymorphism (SNP) database (the 1000 genomes project) [9]. Although these problems have been reported, there has been as yet no attempt to repair the known problems or even change their annotations. Indeed the public data warehouses may be so large that complete data fixup or cleansing cannot be contemplated [10].

Microbiologists are totally familiar with contamination in their wetware experiments but trust their computers. Therefore the chapter serves an important purpose if it convinces practitioners of third millennium biology to be sceptical of computer data (as they are sceptical of physical measurements and living samples), even if the data have come from globally respected sources, such as the NCBI.

This is not just a local problem for a laboratory in Beijing or Tokyo, or even a national problem, but the E-contamination has travelled across national boarders and been actively spread across the globe by international Bioinformatics cooperations in a remarkably short time.

In Section 8 we report an unexpected upside. In the next section (2) in order to make the point that *in vitro* contamination is well known, we summarise HeLa contamination of human cell lines. Section 3 describes the discovery of *in silico* contamination. Whilst Section 4 reports global data contamination takes just a weekend. (The example [8] comes from Christmas 2005). Section 5 reports human (electronic) genes also misbehave. Section 6 summarises Craig Venter's use of computers to construct an artificial species of Mycoplasma, and Section 7 reports Mycoplasma infection of wetware lab samples can occur even in a prestigious $120 million international collaborative project.

2 “HeLa” the Universal Microbiology Human

“HeLa” is the name given to the first successful attempt to grow human cells outside of the body as a cell line [11]. Although derived from a malignant cervix cancer tumour in a single individual, it has been widely used as a model for all human cells. Our interest here is as another example of contamination. Although other human cell lines have since been cultured, poor wet lab practise, lead to many of the available supposedly non-HeLa human cell lines being contaminated with HeLa. Since HeLa grows vigorously, once another cell line is contaminated, natural selection can lead to HeLa dominating.

3 Discovery of Mycoplasma Genes in the Human Genome

Like all the best science stories there is an element of mystery and chance in the discovery of E-contamination. About ten years ago by the river Thames in King’s College, London a triplicated trial was being conducted. For each triple there was a plate for the sample and another for the control. The then state of the art Affymetrix HG-U133 Plus 2.0 GeneChip were used to simultaneously measure the expression of all human genes. (The HG-U133 Plus 2.0 was the first to contain probes for all human genes on a single GeneChip.) The messenger RNA (mRNA) expressed by genes in the first treatment-control pair was measured using a HG-U133_Plus_2 for each and the relative expression of each gene calculated and sorted by their ratio. The largest ratio was about 600 fold difference between treatment and control. In these types of experiments this is a huge difference. What would the next pair give? The second replication gave the same gene, with a ratio of 200. Ok not quite as impressive but still very good. The third replicant? Hmm nothing. But nevertheless two cases, so what was the top gene?

The GeneChip probeset 1570561_at was the one with the large ratio. The Affymetrix HG-U133 Plus 2.0 documentation said it was from a single human expressed sequence tag (EST) (GenBank accession no. AF241217). However GenBank just said (and still says) “Homo sapiens unknown sequence”. Hmm so what to do?

GenBank gives the 176 DNA base sequence for AF241217 and this was queried with BLAST [12] which returned an ordered list of the top 50 matches against every known DNA sequence at the time. The top of the list was naturally AF241217 itself (Homo sapiens unknown sequence) but every other match was for a species of Mycoplasma, leading to the sad conclusion that the initial excitement was misplaced and the impressive gene expression ratio seen was simply because two of the six plates had been contaminated with Mycoplasma.

Rather than just give up and re-run their trial again, the group at King’s continued their investigation in what had happen *in silico*. It turns out (see Figure 2) when Affymetrix in California had come to design the HG-U133 Plus 2.0, they had taken all the “Human” genes from the NCBI’s reference human genome (plus Human DNA sequences from other public databases) and so had included AF241217. Therefore, thinking it was a human gene, they designed their 1570561_at probeset to detect a Mycoplasma gene (AF241217). And true enough years later in the UK it does just that in the King’s College, London laboratory.

Notice although each step has been taken in good faith, a bacterium has infected a human tissue sample, got at least one of its genes sequenced along with human genes and so entered the reference human genome as a "human" gene and stayed there. Years later it is copied from Washington (NCBI) to Silicon Valley (Affymetrix) and transcribed from computer data into physical hardware (GeneChip). Affymetrix have sold many thousands HG-U133 Plus 2.0, globally spreading the fortunate Mycoplasma gene not just *in silico* but now in a novel physical form too. Again years pass, and two HG-U133 Plus 2.0 GeneChips do indeed detect Mycoplasma in London.

Sometime later we discovered a second Mycoplasma gene hiding in the reference human genome [8] (see next section). Both "human" genes [4,8] are lodged in the reference human genome maintained by the NCBI in Bethesda. They have been propagated from the NCBI to the European Bioinformatics Institute (EBI) in Cambridge and other Bioinformatics centres.

Despite repeated acknowledged reports, the NCBI curators refuse to remove the "Human" Mycoplasma DNA sequences or update their annotations. The NCBI argument is that they are simply looking after the data on behalf of the Bioinformatics community. That is, it is up to the owners of the data to say if they want data removed or annotations changed. However it appears whoever uploaded the Mycoplasma sequences thinking they were human are long gone and will never ask the NCBI to repair their error.

Ten years on we can still repeat the BLAST queries made at King's and the results are substantially the same. We still get the same first match and it still claims to be Human and the following matches are for various species of Mycoplasma. However, the exponential increase in DNA sequences in NCBI has continued and it now includes, not just reference species, but also patents. For example, one match is Gen-Probe (of San Diego)'s USA patent 9 212 397, "Compositions and methods for detecting nucleic acid from mollicutes", Filed: June 23, 2010. (Wikipedia says "The best-known genus in the Mollicutes is Mycoplasma".)

4 Speed of Global Mycoplasma Contamination

About five O'Clock Friday afternoon 23 December 2005, 5000 human DNA sequences were uploaded to DDBJ (the DNA Data Bank of Japan). The DNA sequences in DDBJ are mirrored by the NCBI in Washington, so overnight these sequences were automatically transferred across the Pacific to the USA. In turn DNA sequences hosted by NCBI are mirrored by the EBI in Cambridge and other Bioinformatics centres across the globe. Therefore the next night the 5000 DNA sequences were copied across the Atlantic to the EBI in England. However (at least) one of the human DNA sequences was not human but instead was another DNA sequence from a Mycoplasma.

Notice the Mycoplasma sequence has succeeded in spreading itself globally (using the Internet as an unwitting carrier) within 48 hours. Not only did it spread at electronic speed but it has remained lodged in these Bioinformatics data centres labelled as human ever since [8].

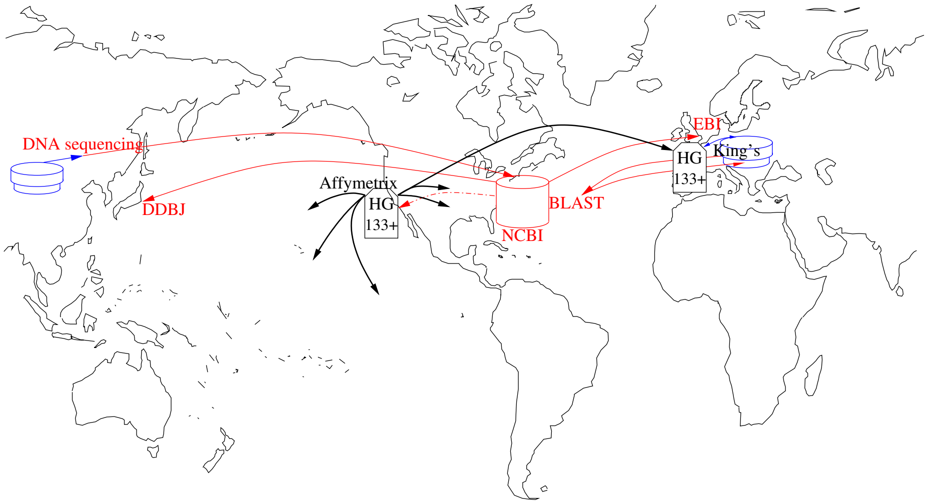

Fig. 2. A path of Mycoplasma *in silico* (red) contamination, Section 3. Mycoplasma infects cell tissue *in vitro* (blue). The culture is sequenced and uploaded (indirectly) into NCBI. From there the gene sequence is distributed globally (e.g. to the DNA Data Bank of Japan and the EBI). Affymetrix also take a copy (dotted red) and include it in their HG-U133 Plus 2.0 GeneChip, which is sold globally (black arrows), including to King's College, London. The Mycoplasma infiltration of the global Bioinformatics infra-structure is discovered when King's run NCBI's BLAST.

5 Inadvertent Human Contamination of Many E-Genomes

Nor is the trade in DNA sequences in just one direction. Mark S. Longo et al. [13] found human DNA sequences all across databases supposedly containing genomes from across most phyla. This suggests poor lab practise and zero E-hygiene is wide spread.

6 Deliberate Human Contamination of *in vivo* Mycoplasma

The to-and-fro of DNA sequences need not be accidental. In 2008 Craig Venter [14] replaced the complete genome of Mycoplasma genitalium with artificial DNA designed by Homo sapiens (with silicon support from digital computers). The DNA was entirely chemically constructed. The artificial Mycoplasma laboratorium grew naturally (albeit in its birth place, a microbiology laboratory). Mycoplasma genitalium having been chosen, in part, as it has one of the smallest genomes of a living organism.

7 Mycoplasma Contamination in 1000 Genomes Project

While the original human genome project mapped the complete human DNA sequences, The 1000 genomes project (1KGP) built on this and mapped the variation in human genetic make up by completely sequencing the genomes of more than a thousand individuals taken from across the world. Naturally the project put their data on the Internet. In [9,15] we analyses these data and concluded at least 7% of their samples were contaminated with Mycoplasma.

This analysis (i.e. [9]) is also potentially useful in alerting users of raw sequence data to that fact that it may not contain what it says it contains, and the need to test assumptions even if multiple factors may interact [16]. Also it may help in identifying microbiology labs with particular problems with Mycoplasma contamination. (As Figure 3 shows, rates of Mycoplasma genes are not uniform in the 1KGP data. E.g. next generation sequence data from the Broad Institute contains proportionately fewer Mycoplasma sequences than those sequenced at WUSTL.)

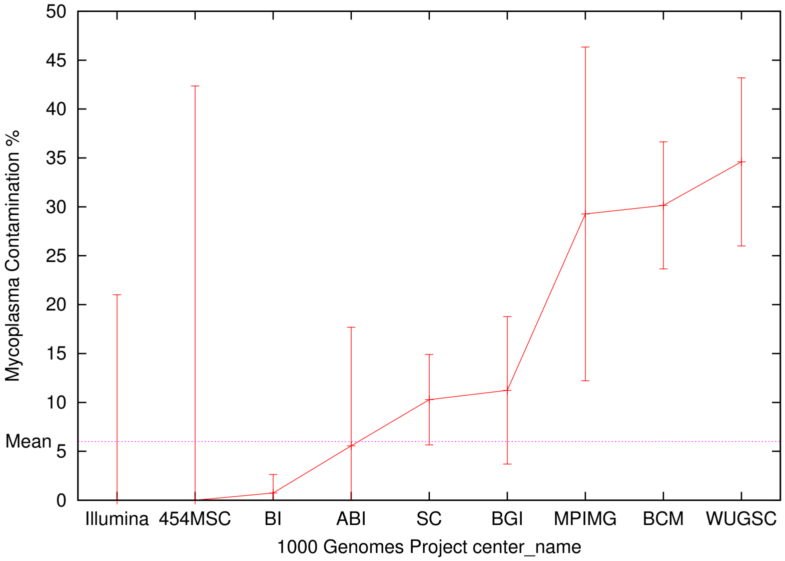

Fig. 3. Fraction of 1000 Genome Project NextGen DNA scans contaminated with Mycoplasma (total average 7%). The Broad Institute (BI) analysed the most samples (90 097) and had little Mycoplasma contamination. Vertical bars indicate estimated error (due to sampling). (Contamination reported in http://www.cs.ucl.ac.uk/staff/W.Langdon/mycoplasma_1000/center_name.html)

8 An Unexpected Benefit to Medicine

Although directly sequencing RNA (RNA-Seq or WTSS) is increasingly popular, it is still common to use Affymetrix GeneChips in gene expression experiments (see Figure 4). As mentioned in Section 3, Affymetrix HG-U133 +2 GeneChips where designed to inadvertently include a Mycoplasma gene [4].

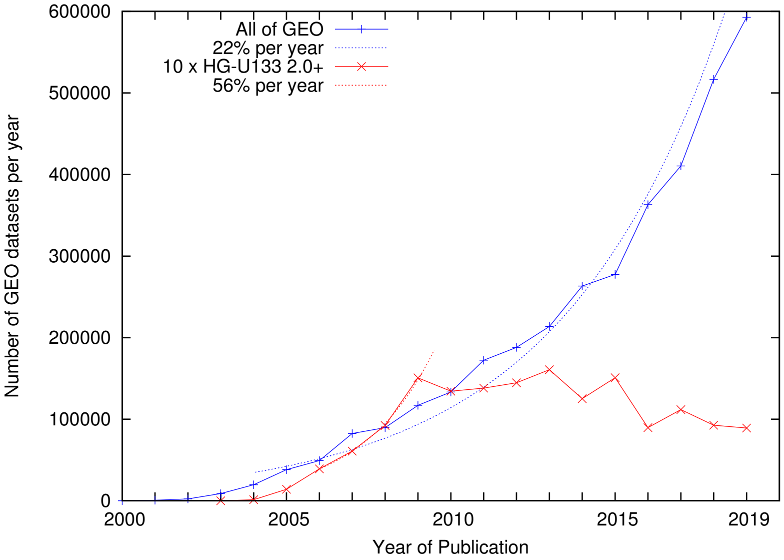

Fig. 4. Data sets in GEO by year of publication. To fit on same vertical scale the values for Affymetrix HG-U133 Plus 2.0 GeneChip multiplied by ten. Notice there may be a lag between running the experiment and publication and sometimes data figure in more than one publication. Dotted lines show exponential per annum growth.

A convention widely enforced by journals has required authors of studies using microarrays to make their data available. Many authors chose to do this by uploading their Affymetrix data to NCBI's Gene Expression Omnibus (GEO) [17]. So now the output of many thousands of HG-U133 Plus 2.0 are archived in GEO. The numbers continue to grow. (Today, 26 Aug 2019, GEO contains more than 150 000, see also Figure 4.) In 2007 we down loaded all the HG-U133 +2 data in GEO. After cleaning and removing duplicates we had 2757 samples covering a wide variety of human tissue types and disease states, from numerous individuals. These were added to the University of Essex's RNAnet [18,19]. 2757 Affymetrix CEL files was far more than any individual laboratory or even consortium could assemble and so RNAnet enabled quantile normalisation [20] across a huge sample. RNAnet provided ready access to the normalised results of *in vitro* gene expression experiments run across the globe and showed the previously unknown presence to Mycoplasma contamination in the labs in five published results. Contacting the authors of the five publications produced a variety of positive responses in all cases [21,22,23,24,25,26,27]. Notice by exploiting the power of the existing Big Data Bioinformatics infrastructure we can find Mycoplasma infection globally by analysing existing data without the use of special equipment or patented inventions (cf. Section 3 above).

9 Summary: Be Aware of *In Silico* Data Contamination

The problem of Mycoplasma contamination in microbiology laboratories has become well known and reputable laboratories take the problem very seriously. Nonetheless Mycoplasma contamination in wet labs has resulted in indirect contamination of globally interlinked computer data banks. (These databases can in many ways rightly be regarded as a triumph of E-Science.) However their corruption has not been treated with dispatch. In other words Biologists are well aware of *in vitro* contamination but many are unaware of *in silico* contamination and so treat data in a computer as gospel.

References

1. Mendel, G.: Experiments in plant hybrizization. Verhandlungen des naturforschenden Vereines in Brno (IV), 3–47 (1865), `http://www.esp.org/foundations/genetics/classical/gm-65.pdf`, translated by William Bateson in 1901 (updated Roger Blumberg, etc.)
2. McClintock, B.: A cytological and genetical study of triploid maize. Genetics 14(2), 180–222 (1929), `http://www.genetics.org/content/14/2/180.short`
3. Akiba, T., Koyama, K., Ishiki, Y., Kimura, S., Fukushima, T.: On the mechanism of the development of multiple-drug-resistant clones of Shigella. Japanese Journal of Microbiology 4, 219–227 (Apr 1960), `http://dx.doi.org/doi:10.1111/j.1348-0421.1960.tb00170.x`
4. Aldecoa-Otalora, E., Langdon, W.B., Cunningham, P., Arno, M.J.: Unexpected presence of mycoplasma probes on human microarrays. BioTechniques 47(6), 1013–1016 (December 2009), `http://dx.doi.org/doi:10.2144/000113271`
5. Nakabachi, A., Yamashita, A., Toh, H., Ishikawa, H., Dunbar, H.E., Moran, N.A., Hattori, M.: The 160-kilobase genome of the bacterial endosymbiont carsonella. Science 314(5797), 267 (13 October 2006), `http://dx.doi.org/doi:10.1126/science.1134196`
6. Drexler, H.G., Uphoff, C.C.: Mycoplasma contamination of cell cultures: Incidence, sources, effects, detection, elimination, prevention. Cytotechnology 39(2), 75–90 (2002), `http://dx.doi.org/doi:10.1023/A:1022913015916`
7. Miller, C.J., Kassem, H.S., Pepper, S.D., Hey, Y., Ward, T.H., Margison, G.P.: Mycoplasma infection significantly alters microarray gene expression profiles. BioTechniques 35(4), 812–814 (October 2003), `http://dx.doi.org/doi:10.2144/03354mt02`
8. Langdon, W.B., Arno, M.J.: More mouldy data: Another mycoplasma gene jumps the silicon barrier into the human genome. ArXiv e-prints (14 June 2011), `http://arxiv.org/abs/1106.4192`

9. Langdon, W.B.: Mycoplasma contamination in the 1000 genomes project. BioData Mining 7(3) (29 April 2014), http://dx.doi.org/doi:10.1186/1756-0381-7-3
10. Altenhoff, A.M., Glover, N.M., Train, C.M., Kaleb, K., Warwick Vesztrocy, A., Dylus, D., de Farias, T.M., Zile, K., Stevenson, C., Long, J., Redestig, H., Gonnet, G.H., Dessimoz, C.: The OMA orthology database in 2018: retrieving evolutionary relationships among all domains of life through richer web and programmatic interfaces. Nucleic Acids Research 46(D1), D477–D485 (4 January 2018), https://doi.org/10.1093/nar/gkx1019
11. Masters, J.R.: HeLa cells 50 years on: the good, the bad and the ugly. Nature Reviews Cancer 2(4), 315–319 (April 2002), http://dx.doi.org/doi:10.1038/nrc775
12. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J.: Gapped BLAST and PSI-BLAST a new generation of protein database search programs. Nucleic Acids Research 25(17), 3389–3402 (1997), http://dx.doi.org/doi:10.1093/nar/25.17.3389
13. Longo, M.S., O'Neill, M.J., O'Neill, R.J.: Abundant human DNA contamination identified in non-primate genome databases. PLoS ONE 6(2), e16410 (02 2011), http://dx.doi.org/10.1371%2Fjournal.pone.0016410
14. Gibson, D.G., Benders, G.A., Andrews-Pfannkoch, C., Denisova, E.A., Baden-Tillson, H., Zaveri, J., Stockwell, T.B., Brownley, A., Thomas, D.W., Algire, M.A., Merryman, C., Young, L., Noskov, V.N., Glass, J.I., Venter, J.C., Hutchison, C.A., Smith, H.O.: Complete chemical synthesis, assembly, and cloning of a mycoplasma genitalium genome. Science 319(5867), 1215–1220 (2008), http://dx.doi.org/doi:10.1126/science.1151721
15. Grens, K.: Mistaken identities. The Scientist (Jan 1 2015), https://www.the-scientist.com/news-opinion/mistaken-identities-36160
16. Petke, J.: Constraints: The future of combinatorial interaction testing. In: 2015 IEEE/ACM 8th International Workshop on Search-Based Software Testing. pp. 17–18. Florence (May 2015), http://dx.doi.org/doi:10.1109/SBST.2015.11
17. Barrett, T., Troup, D.B., Wilhite, S.E., Ledoux, P., Rudnev, D., Evangelista, C., Kim, I.F., Soboleva, A., Tomashevsky, M., Edgar, R.: NCBI GEO: mining tens of millions of expression profiles–database and tools update. Nucleic Acids Research 35(Database issue), D760–D765 (January 2007), http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=pubmed&dopt=Abstract&list_uids=17099226
18. Langdon, W.B., Harrison, A.P., Sanchez Graillet, O.: RNAnet a map of human gene expression. In: EMBO-2008. Heidelberg (15-18 Nov 2008), http://www.cs.ucl.ac.uk/staff/W.Langdon/ftp/papers/RNAnet-EMBO2008.pdf, abstract presented
19. Sanchez-Graillet, O., Stalteri, M.A., Rowsell, J., Upton, G.J., Harrison, A.P.: Using surveys of Affymetrix GeneChips to study antisense expression. Journal of Integrative Bioinformatics 7(2), 114 (2010), http://dx.doi.org/doi:10.2390/biecoll-jib-2010-114
20. Bolstad, B.M., Irizarry, R.A., Åstrand, M., Speed, T.P.: A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 19(2), 185–193 (February 2003), https://doi.org/10.1093/bioinformatics/19.2.185
21. Langdon, W.B., Arno, M.: *In Silico* infection of the human genome. In: Giacobini, M., Vanneschi, L., Bush, W.S. (eds.) 10th European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics, EvoBIO 2012. LNCS, vol. 7246, pp. 245–249. Springer Verlag, Malaga, Spain (11-13 April 2012), http://dx.doi.org/doi:10.1007/978-3-642-29066-4_22

22. Langdon, W.B.: Correlation of microarray probes give evidence for mycoplasma contamination in human studies. In: Smith, S.L., Cagnoni, S., Patton, R.M. (eds.) GECCO-2013 Workshop: MedGEC Medical Applications of Genetic and Evolutionary Computation. pp. 1447–1454. ACM, Amsterdam (6-10 July 2013), http://doi.acm.org/10.1145/2464576.2482725
23. El Hader, C., Tremblay, S., Solban, N., Gingras, D., Beliveau, R., Orlov, S.N., Hamet, P., Tremblay, J.: HCaRG increases renal cell migration by a TGF-alpha autocrine loop mechanism. Am J Physiol Renal Physiol 289(6), F1273–F1280 (Dec 2005), http://dx.doi.org/doi:10.1152/ajprenal.00103.2005
24. Schmidt, S., Rainer, J., Riml, S., Ploner, C., Jesacher, S., Achmueller, C., Presul, E., Skvortsov, S., Crazzolara, R., Fiegl, M., Raivio, T., Jaenne, O.A., Geley, S., Meister, B., Kofler, R.: Identification of glucocorticoid-response genes in children with acute lymphoblastic leukemia. Blood 107(5), 2061–2069 (March 1 2006), http://dx.doi.org/doi:10.1182/blood-2005-07-2853
25. Mayburd, A.L., Martlinez, A., Sackett, D., Liu, H., Shih, J., Tauler, J., Avis, I., Mulshine, J.L.: Ingenuity network-assisted transcription profiling: Identification of a new pharmacologic mechanism for MK886. Clin Cancer Res 12(6), 1820–1827 (Mar 15 2006), http://dx.doi.org/doi:10.1158/1078-0432.CCR-05-2149
26. Jack, G.D., Cabrera, M.C., Manning, M.L., Slaughter, S.M., Potts, M., Helm, R.F.: Activated stress response pathways within multicellular aggregates utilize an autocrine component. Cellular Signalling 19(4), 772–781 (2007), http://dx.doi.org/doi:10.1016/j.cellsig.2006.10.005
27. Cappellen, D., Schlange, T., Bauer, M., Maurer, F., Hynes, N.E.: Novel c-MYC target genes mediate differential effects on cell proliferation and migration. EMBO Rep 8(1), 70–76 (Jan 2007), http://dx.doi.org/doi:10.1038/sj.embor.7400849, European Molecular Biology Organization