UCL DEPARTMENT OF
COMPUTER SCIENCE

Research Note

RN/13/10

Mycoplasma Contamination in The 1000 Genomes Project

22 May 2013

W. B. Langdon

Abstract

Mapping next generation DNA sequences from the thousand genome project against published genomes reveals many that match one or more Mycoplasma but are not included in the reference human genome GRCh37.p5. Many of these are of low quality but NCBI BLAST searches confirm some high quality, high entropy sequences match Mycoplasma but no human sequences. Suggesting at least 7% of 1000G samples are contaminated.

1 Introduction

Mycoplasma are tiny bacteria which readily grow in cell culture media. They have small genomes. Contamination of molecular biology laboratories by them is widespread in [1]. Because of their small size etc. they are hard to detect. Depending upon medium, Mycoplasma contamination rates of 1% to 15–35% (or even higher) have been reported [2]. Mycoplasma contamination can render cell line gene expression measurements unreliable [1]. Many labs routinely sterilised their equipment to counter it. About 1% of published NCBI's GEO [3] GeneChip data appear to be contaminated [4]. GEO contains gene expression data, here we start to look for similar contamination in genome studies. The 1000 Genomes Project [5] is an international collaboration which has mapped in whole or in part the genomes or more then 2500 individuals and published studies of SNPs and other human genetic variations. We selected The 1000G Project, since it investigates human genetic material, is widely respected, it covers many sites with diverse data sources and has made available vast quantities of its raw data.

2 Method

The master index file, `sequence.index`, which describes all the current 1000 genomes project datasets was down loaded[1]. As of 8 February 2013 there were 47,315 datasets available (a further 208 had been withdrawn). The available datasets comprised: 39 736 paired-end and 4822 single ended DNA sequence measurements plus a further 1611 (paired end) and 938 (single ended) datasets which used ABI_SOLID colorspace encoding. Figure 1 shows the distribution of DNA sequence lengths. Almost all colorspace sequences contain 25, 35 or 50 base pairs, whereas lengths 68, 76, 100 and 101 dominate non-colorspace sequences. 4058 datasets were randomly chosen and down loaded. All the DNA measurements are in fastq format, so they include a quality score per DNA base pair. As is to be expected, most of the measurements were paired double ended sequences but about 11% consist of only one DNA sequence. Also most data are conventionally coded but about 5% are SOLiD colorspace encoded. Each dataset contains DNA sequences of the same length. Typically sequences are 26–108 base pairs (approx 100bp on average).

On average: each datasets contained 13 million DNA sequences (or pairs of sequences). Even compressed, each file is approximately a gigabyte. (Compression reduces download size by a factor of about 3.1) Paired end datasets need two such files. The down load speed was variable, typically between $2.5\ 10^6$ and $36\ 10^6$ bytes/second, with a mean of 11 million bytes per second. In total 7547 files were down loaded (6.0 terabytes) containing 51 494 393 834 DNA measurements totalling about $7.5\ 10^{12}$ base pairs.

We then used Bowtie [6] to find those DNA measurements (i.e. DNA sequences or pairs of DNA sequences) which matched one or more of the published Mycoplasma genomes but do not match the reference human genome GRCh37.p5. See Figure 2. We used all of the Mycoplasma genomes available from NCBI (30 in total, see [7, Appendix A]). Apart from using multiple threads `-p8`, Bowtie's defaults were used through out. The Bowtie EBWT databases for the normal and colorspace Mycoplasma genomes are both 36 MBytes. Despite including 30 species, due to the small size of Mycoplasma genomes, they are both considerably smaller than the two databases for the human reference genome, which are 2.9 GB for both normal and colorspace. The Bowtie EBWT databases and colorspace databases for the human reference genome GRCh37.p5 include all sequences. I.e., as well as chromosomal DNA, they both include human mitochondrial, "unlocalized" and "unplaced" sequences.

Although the server was shared with other users, Figure 2 gives a reasonable indication of Bowtie's performance. Notice Bowtie is usually faster on single ended rather than paired double ended DNA sequences (mean 28 v. 18 million sequences per hour per CPU). Although downloading and decompressing the files took 37% of the elapsed time, despite using all 8 CPU cores, almost all the remaining 63% of time was used by Bowtie.

[1] ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/sequence.index

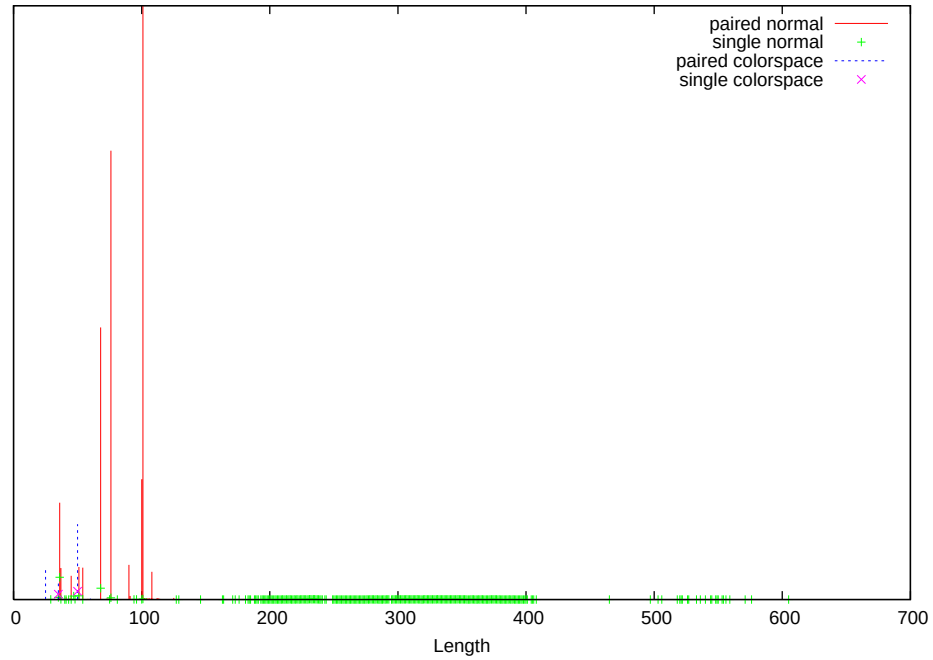

Figure 1: Lengths of DNA sequences for Thousand Genomes Project. Mostly measurements have two paired ends. The mode is for each end to have 101 DNA base pairs.

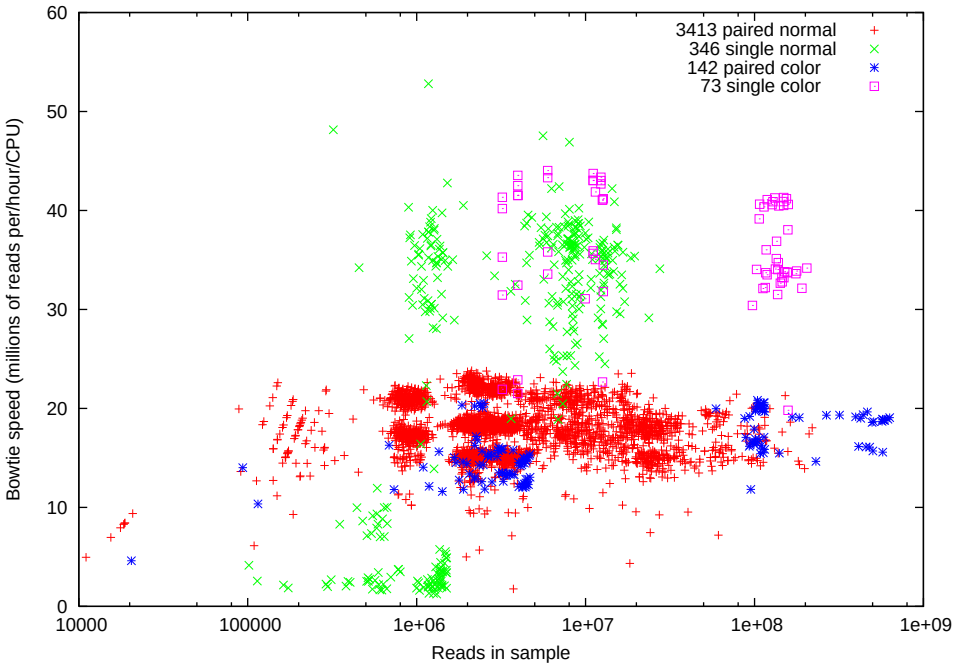

Figure 2: Speed of Bowtie mapping short nextGen Thousand Genomes Project DNA sequences against 30 Mycoplasma genomes. (32 GB 8 core 3 GHz server. Note log horizontal scale.)

Table 1: Thousand Genomes Project nextGen DNA which match Mycoplasma (with on average three or fewer mismatches) but not the reference human genome GRCh37.p5

Type		Measurements	Datasets	Fraction of datasets
pair	ordinary	797	106	3%
pair	colorspace	17015	111	77%
single	ordinary	752	108	28%
single	colorspace	4020	72	96%
	Totals	22584	397	10%

3 Results

Bowtie found 4 803 930 DNA measurements which match one or more Mycoplasma genomes. Almost all these also matched somewhere in the reference human genome, leaving 75 879 which match Mycoplasma but do not appear to be human. These occur in 2055 of our 3983 samples (51.6%).

Since nextGen sequences are noisy it is to be expected that a large number will fail to match in the human reference genome. (We measure about 30%). However some of these unmatched DNA measurements may not be simply due to noise. These are the ones we investigate to see if they could be due to Mycoplasma contamination.

3.1 Number of Mismatches between 1000 Genome Project DNA and Mycoplasma

Figure 3 shows that although Bowtie finds matches for each of these 75 879 against the Mycoplasma genomes, the exactness of the match varies considerably, from 100% matches to 78 mismatches. Figure 3 also breaks these figures down into pair end and single DNA strands and Solexa coding type (normal v. SOLiD colorspace). Although the colorspace encoding represents a small fraction of the whole data, of the DNA measurements which match Mycoplasma only and for which Bowtie reports (on average) three or fewer mismatches, 93% of them are colorspace encoded. (Notice however colorspace sequences tend to be much shorter, see Figure 1.) On average, if affected, colorspace datasets contain many more affected DNA measurements than normally coded Solexa datasets. (See columns 3–4 of Table 1.) Overall ten percent of the 1000 Genomes Project datasets contain sequences which match Mycoplasma well (i.e. on average ≤ 3 mismatches) but do not appear in the reference human genome (last figure in Table 1).

3.2 Quality of The 1000 Genomes Project DNA measurements

Solexa data, like that from other nextGen scanners, are inherently noisy. Solexa provides an estimate of the signal to noise ratio (expressed as $-1.0 \log_{10}$) per base position in each DNA sequence. This can easily mount up to several hundred quality values. To stably condense these into a manageable statistic, we ignore the worse and second to worst base in each DNA sequence and use the third worst. For paired end data, we use worst of the two ends.

If we compare the quality of DNA measurements which match Mycoplasma but which do not occur in the reference human genome (Figure 4) with those which do match GRCh37.p5 we see in both cases measurements with a large numbers of mismatches only occur in low quality data. (Figure 5 reports a typical run.) Further Figure 4 makes it plain that most of the DNA measurements which match Mycoplasma but which do not occur in the reference human genome contain at least three poor quality values. There are 1944 measurements with a quality above 0.5 (which match one or more Mycoplasma genomes with ≤ 3 mismatches). They occur in 269 datasets. (7% of the sample. See last number in Table 2).

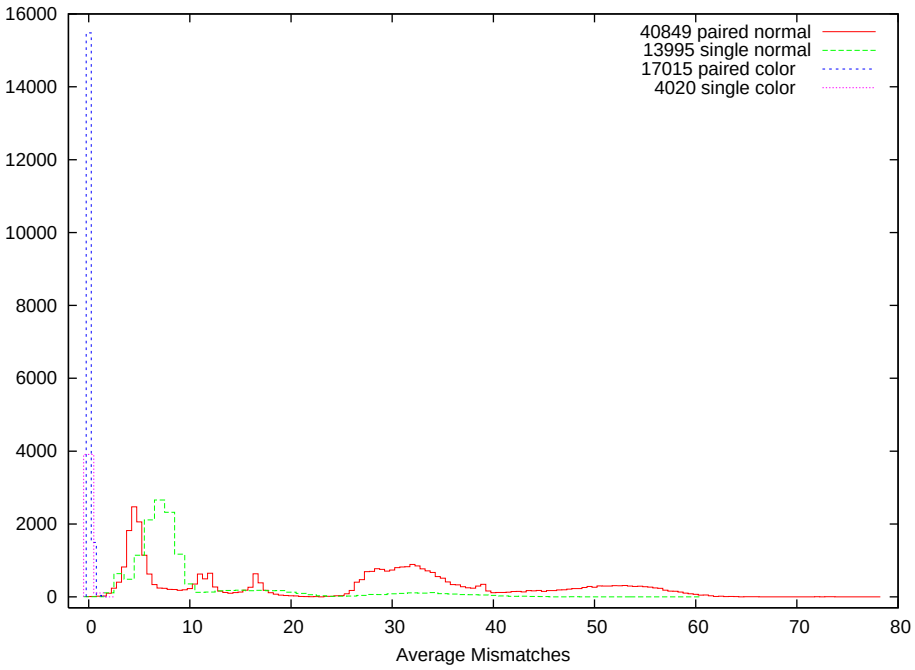

Figure 3: Distribution of mismatches in matches against Mycoplasma genomes for 75 879 Thousand Genomes Project DNA sequences which not match the reference human genome.

Table 2: High quality Thousand Genome Project nextGen DNA which match Mycoplasma but not the reference human genome GRCh37.p5 (NB. no more than three bases with quality worse than 0.5 and on average three or fewer mismatches.)

Type		Measurements	Datasets	Fraction of datasets
pair	ordinary	542	87	3%
pair	colorspace	1042	63	43%
single	ordinary	234	78	20%
single	colorspace	126	41	55%
	Totals	1944	269	7%

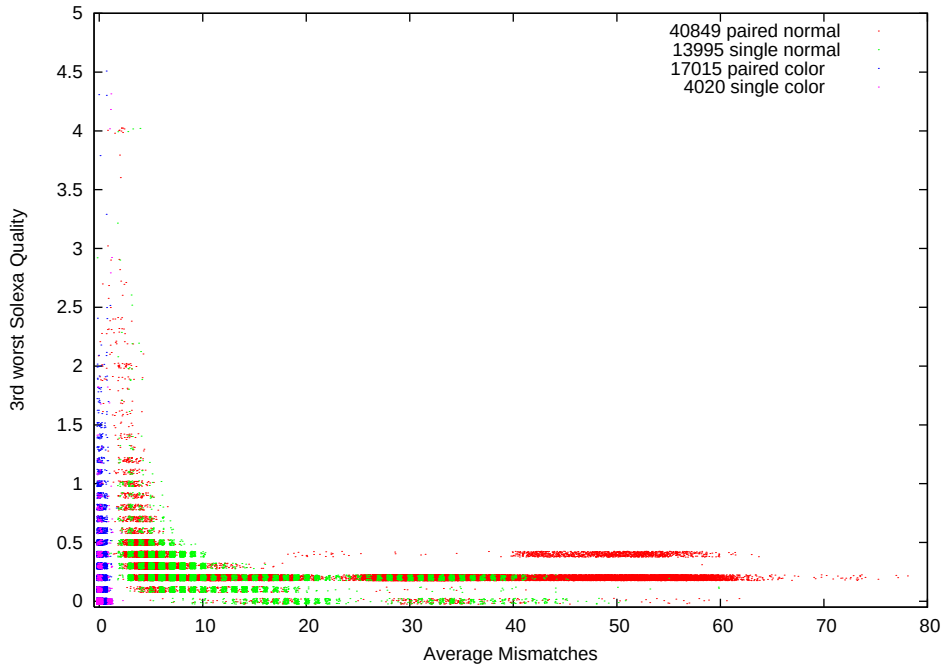

Figure 4: Quality of 75 879 Thousand Genomes Project sequences which match one or more Mycoplasma genomes but do not match the reference human genome. Above 1.3 is a p-value better than 0.05. Horizontal and vertical noise added to spread data.)

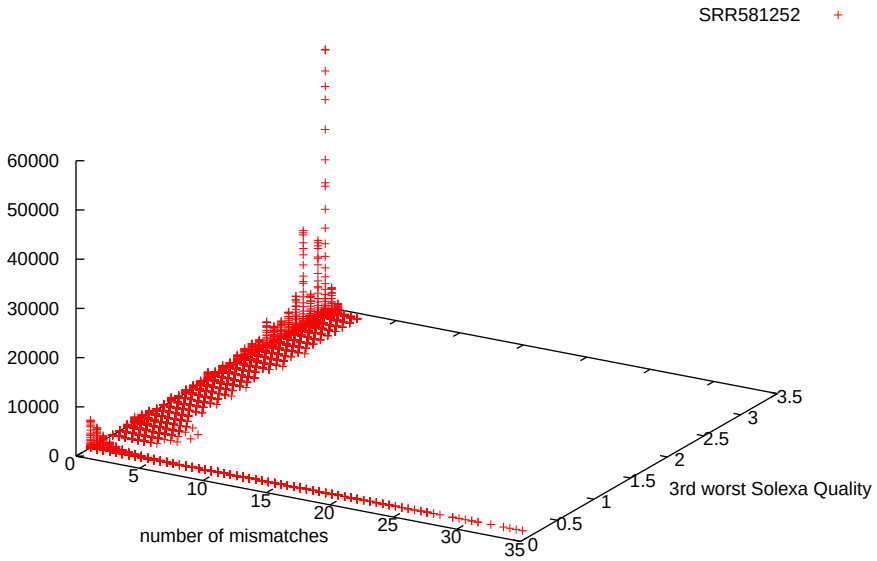

Figure 5: Quality of 1 762 302 DNA pairs which match the human reference genome. (From an example 1000 Genomes Project paired-end run.) Showing typically large numbers of mismatches are only reported for poor quality data.

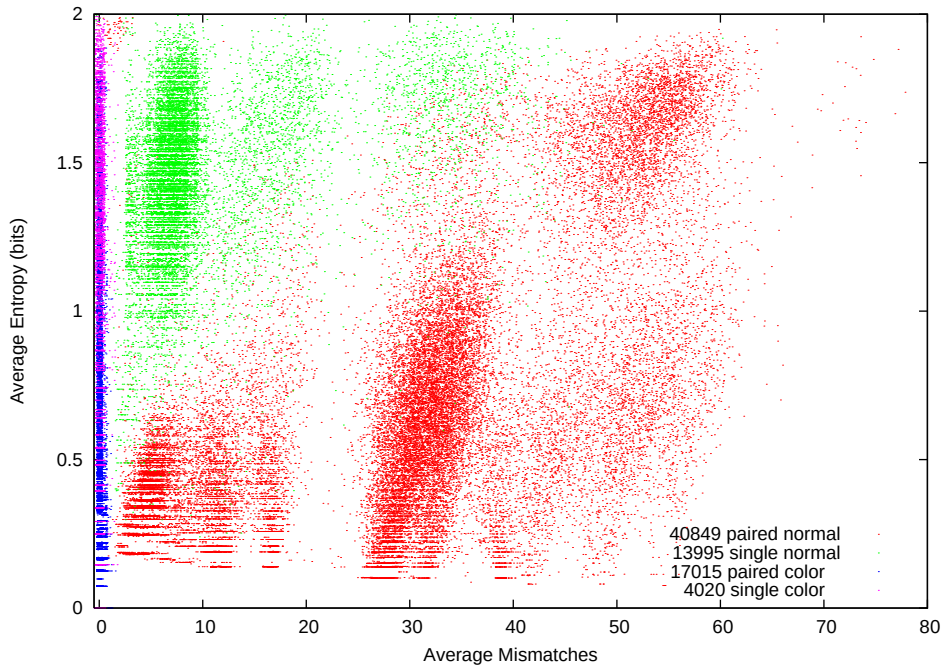

Figure 6: Entropy per DNA base of 75 879 Thousand Genomes Project sequences which match one or more Mycoplasma genomes but do not match the reference human genome. (See also Figure 7. Horizontal noise added to spread data.)

3.3 Entropy of 1000 Genomes Project DNA matching Mycoplasma

Figure 6 shows that the exactness with which the DNA measurements match Mycoplasma and the entropy (incompressibility) of its sequences appears to be unrelated. For the very much larger volume of sequences which do match the human reference genome, entropy also plays little role. Instead large numbers of mismatches occur only in low entropy sequences. (Figure 8 plots data from a typical 1000 Genomes Project run.) Although Bowtie reports a match, in some cases Bowtie must change many (up to 78) individual DNA bases to get an exact match between the measured DNA sequences and one of the published Mycoplasma genomes. Low entropy (compressible) DNA sequences are highly repetitive. Many real genomes have highly repetitive regions. A highly repetitive simple DNA pattern (even if it exactly matches against a genome) is liable to fall in repetitive region of a (published) genome, where coverage is liable to be patchy. See also Figure 7, which concentrates on Mycoplasma only DNA measurements which match Mycoplasma genomes well.

3.4 Confirming Bowtie with NCBI BLAST

Rather than trying to run BLAST on several thousand DNA strings, we added entropy, a higher quality threshold and exact matching, to choose the best sequences and then ran BLAST on these. In detail, we used a quality threshold above 1.3, we ignored repetitive DNA sequences (i.e. average entropy below 1.0) and requiring at least one exact match against one of our Mycoplasma genomes. This gives seven measurements (none of which is from a SOLiD colour space scanner). See Table 3 on page 9. BLAST provides strong evidence that these DNA measurements are really from one or more Mycoplasma species.

We used NCBI's Blast [9] program to confirm our Bowtie results. (We used the default parameters provided by the EBI web interface except we request the first 1000 matches, rather than the first 50 matches.) Using BLAST on each of the sequences in Table 3 shows each of the seven high quality DNA measurements do, as expected, match one or more species of Mycoplasma and none matches the reference human genome. In a few cases the second pair matches "Homo sapiens clones", rather than the human reference sequence. Often these are draft sequences and only in one case (ERR013159.14600701) do both ends of DNA pair match the clone.

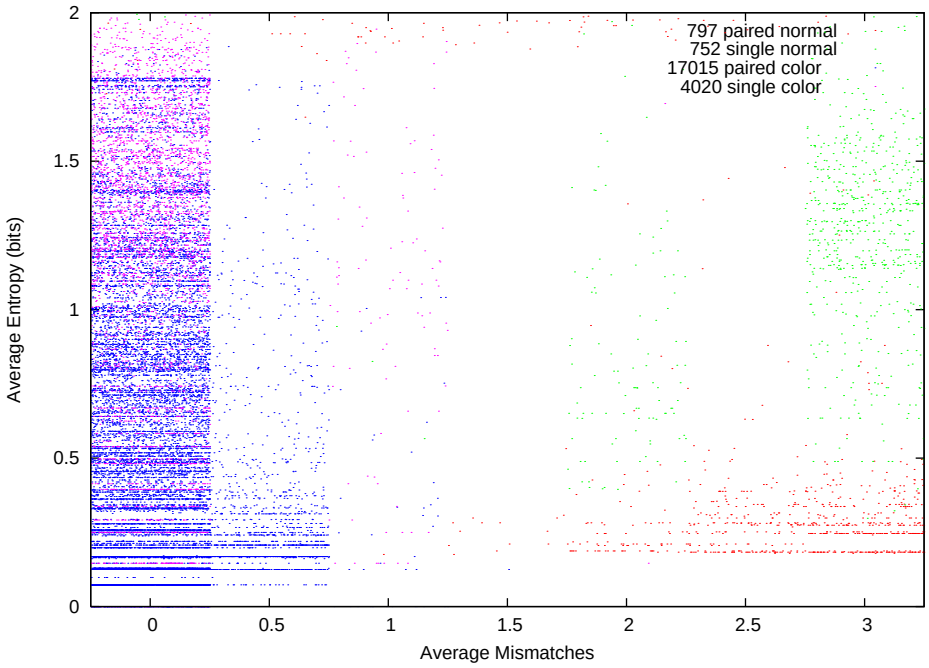

Figure 7: Entropy per DNA base of 22 584 Thousand Genomes Project sequences which match one or more Mycoplasma genomes with 3 or fewer mismatches but do not match the reference human genome. (Detail of Figure 6. Horizontal noise added to spread data.)

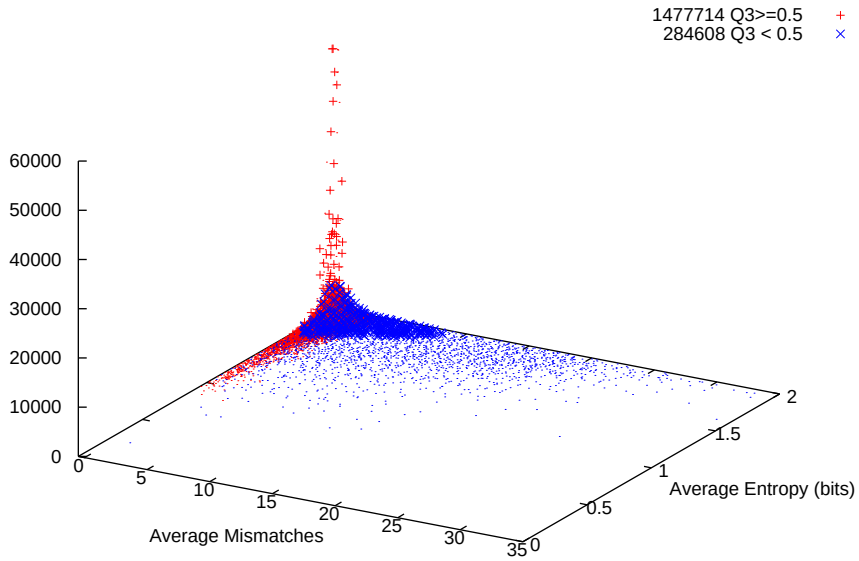

Figure 8: Entropy v. number of mismatches for 1 762 302 DNA pairs which match the human reference genome. (From the same example 1000G paired-end run as in Figure 5.) Most DNA measurements which match GRCh37.p5 are not repetitive (i.e. have high entropy). Also low quality ($\times$) measurements tend to have more mismatches.

The final column of Table 3 reports an example of one of the Mycoplasma genes which BLAST finds which match the DNA sequence. In the case of paired end DNA measurements, BLAST has been run separately on both end. The reported gene is matched by both ends. (In three cases an example gene has not been chosen because BLAST matches the whole of, a number of, Mycoplasma genomes.) Noting the example gene's similarity, it is tempting to ascribe some biological meaning to the gene, however BLAST effectively searches all the published DNA sequences and so the similarity may well simply reflect a bias in the published sequences. Ribosomal DNA is highly conserved and has been heavily studied as a tree of life phylogenetic marker of evolutionary inheritance. Making it one of the more frequent genes in EMBL-Bank.

We take BLAST's matches and the lack of BLAST matches against the official human reference genome as confirming our Bowtie results. That is, Table 3 suggests samples ERR009050, ERR002459, ERR013159 and ERR022473 appear to have been contaminated with Mycoplasma. However, of these four, only in one (ERR009050) are there more than a few score DNA measurements which Bowtie matches against Mycoplasma.

4 Discussion

Mycoplasma contamination in microbiological laboratories is wide spread [2]. Previously we have reported Mycoplasma contamination in gene expression data from a number of different laboratories on three continents [4; 7; 10]. Indeed wet lab contamination is so wide spread that Mycoplasma genes have managed to jump the silicon barrier and get themselves up loaded into international data banks [11]. It appears about 1% of published gene expression data are affected [4]. Here we have analysed DNA sequences directly, rather than gene expression. While the techniques are totally different, there is still scope for sample contamination and sequence comparison (Section 3) suggests at least 7% of public data provided by the 1000 Genomes Project may have some Mycoplasma contamination. However the fraction may be higher due to: overlap in DNA sequence space between Human and Mycoplasma genomes and due to excluding low quality data.

Whilst the problem of contamination of nextGen sequences has been considered before, previous studies, e.g. Jun et al. [12] and Cibulskis et al. [13] have looked at contamination by other member of the same species. Indeed there have been several reports of unexpected personal (i.e. human) DNA in The Thousand Genomes Project public data but no reports of non-human contamination. However we downloaded and scanned a random sample of more than 50 billion DNA measurements from their FTP site and found tens of thousands which may have come from Mycoplasma contamination.

Table 3: High quality, non-repetitive 1000 Genomes Project DNA measurements which match one or more published Mycoplasma genomes but which do not match the reference human genome. Entropy per bit (column 1), 3rd worst quality (column 2), file and sequence id (column 3) of Solexa DNA strands (column 4). Column 5 gives an example gene (see Section 3.4).

2.0	2.9	ERR009050.2605525	GCCGTAACTATAACGGTCCTAAGGTAGCGAAATTCCTTGTC	S16 23S ribosomal RNA
2.0	2.3	ERR002459.4464466	ACGGTTTTCAAGACCGTTCCCTTCAGCCAGACTTGG CCTGACGGTTTTCAAGACCGTTCCCTTCAGCCAGAC	transfer RNA-Ser
1.8	2.1	ERR013159.14600701	CGCTTTCATTGTTCCGCCAGTAGCTAAAACATCATCAAT AATTGCTACTTTTTGGCCTTTTTTCAACATATTAGTTTG GATTTCTAGAGTTGATTTACCATATTCTAA TTTTTGGCCTTTTTTCAACATATTAGTTTGGATTTCTAGA GTTGATTTACCATATTCTAAATCATACTCAAAACTAATA ACGTCTCCTGGTAATTTTTTAGGTTTTCT	
1.8	2.0	ERR013159.12593030	GAGCTTGTTTTTCGTATTTTTCAATTTCTATTTCGTCATT GATTTGTCAATTTGGTAAATTTGTGTTTTCGCTATCAGG TTTGGTTAGTTTAAAATAACCATCAAAAG AGGTTTGGTTAGTTTAAAATAACCATCAAAAGTAATTA TTGAACCAGAAAGATAAAATTTGTGTTCTTGATTTAAA AATTCATAACGTGTAATTTGTCTTTCAGGAAC	
1.7	2.2	ERR013159.18901091	GGTCAAGTTTACAACAAAATGTTTGCACTTCAAAAAGA ACTAGAAGAACTAGAAGAAAATAAAGAAGAAAATACT TTAATCAAAGAAGTAGTGAACCAAGAAGATATT AAAAGAACTAGAAGAACTAGAAGAAAATAAAGAAGAA AATACTTTAATCAAAGAAGTAGTGAACCAAGAAGATAT TGCAAATATTGTTTCTAAATGAACAAAAATTCC	
2.0	1.9	ERR013159.7037432	TCTAGAGATACTGCCTGGGTAACCAGGAGGAAGGTGG GGACGACGTCAAATCATCATGCCTCTTACGAGTGGGGC AACACACGTGCTACAATGGTCGGTACAAAGAGA AGTGGGGCAACACACGTGCTACAATGGTCGGTACAAA GAGAAGCAATATGGTGACATGGAGCAAATCTCAAAAA ACCGATCTCAGTTCGGATTGAAGTCTGCAACTCG	16S ribosomal RNA
1.9	1.9	ERR022473.14544768	TGCTTTTTTACCTCATGGAGTAAGTGGTGCTTTACGTCC AATTGGTTGTTTACCTTCACCACCACCATGTGGGTGATC ATTTGGGTTCATTACAGAACCTCTAACTGT GGTGATCATTTGGGTTCATTACAGAACCTCTAACTGTTG GACGAATACCTAAATGACGATTACGTCCTGCTTTTCCA ATGTTAACTAGGTTATGTTCTTCATTTCCTA	ribosomal protein cluster

References

[1] 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.

[2] 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.

[3] Tanya Barrett, Dennis B. Troup, Stephen E. Wilhite, Pierre Ledoux, Dmitry Rudnev, Carlos Evangelista, Irene F. Kim, Alexandra Soboleva, Maxim Tomashevsky, and Ron Edgar, "NCBI GEO: mining tens of millions of expression profiles–database and tools update," *Nucleic Acids Research*, vol. 35, no. Database issue, pp. D760–D765, January 2007.

[4] 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.

[5] Richard M. Durbin, *et al.*, "A map of human genome variation from population-scale sequencing," *Nature*, vol. 467, no. 7319, pp. 1061–1073, 28 Oct 2010.

[6] 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.

[7] 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.

[8] Robert Schmieder and Robert Edwards, "Fast identification and removal of sequence contamination from genomic and metagenomic datasets," *PLoS ONE*, vol. 6, no. 3, 03 2011.

[9] Stephen F. Altschul, Thomas L. Madden, Alejandro A. Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman, "Gapped BLAST and PSI-BLAST a new generation of protein database search programs," *Nucleic Acids Research*, vol. 25, no. 17, pp. 3389–3402, 1997.

[10] W. B. Langdon, "Correlation of microarray probes give evidence for mycoplasma contamination in human studies," in *GECCO-2013 Workshop: MedGEC Medical Applications of Genetic and Evolutionary Computation*, Stephen L. Smith, Stefano Cagnoni, and Robert M. Patton, Eds., Amsterdam, 6-10 July 2013, ACM, To appear.

[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] Goo Jun, Matthew Flickinger, Kurt N. Hetrick, Jane M. Romm, Kimberly F. Doheny, Goncalo R. Abecasis, Michael Boehnke, and Hyun Min Kang, "Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data," *American Journal of Human Genetics*, vol. 91, no. 5, pp. 839–848, Nov 2 2012.

[13] Kristian Cibulskis, Aaron McKenna, Tim Fennell, Eric Banks, Mark DePristo, and Gad Getz, "Contest: estimating cross-contamination of human samples in next-generation sequencing data," *Bioinformatics*, vol. 27, no. 18, pp. 2601–2602, 2011.