Draft BioData Mining, 2014, 7:3, DOI: 10.1186/1756-0381-7-3

Mycoplasma Contamination in The 1000 Genomes Project

W. B. Langdon

Department of Computer Science, University College London, Gower Street, London WC1E 6BT, UK

Email: w.langdon@cs.ucl.ac.uk;

*Corresponding author

Abstract

Background: *In silco* Biology is increasingly important and is often based on public data. While the problem of contamination is well recognised in microbiology labs the corresponding problem of database corruption has received less attention.

Results: Mapping 50 billion 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.

Conclusions: It appears at least 7% of 1000G samples are contaminated.

Keywords

Molecular Biology, Microbiology, genetics, metagenomic, Data mining, Next-generation DNA sequencing, Data cleansing, High Throughput, Solexa, 454, SOLiD.

Background

Mycoplasma are tiny bacteria which readily grow in cell culture media. They have small genomes. Contamination of molecular biology laboratories by them is widespread [1]. Their small size makes them 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 Gene Expression Omnibus (GEO) [3] GeneChip data appear to be contaminated [4] [5]. Indeed wet lab contamination is so wide spread that Mycoplasma genes have managed to jump the silicon barrier and get themselves incorporated into international data banks as *Human* genes [6].

GEO contains gene expression data, here we start to look for similar contamination in genome studies. The 1000 Genomes Project [7] 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 1000 Genomes 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.

Results and Discussion

Bowtie (version 0.12.7) [8] found 4 803 930 DNA measurements in a random sample which match one or more Mycoplasma genomes (see Figure 1)[1]. 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 are non-uniformly clustered in 51.6% of individual DNA samples.

NextGen scanners are noisy. So, on the assumption that errors are independent, typically multiple (e.g. 3) scans are run. However non-uniform clusters of errors indicate that they are not independent and therefore redundant scans may not resolve the problem. Noise may be part of the reason why Bowtie reports about 30% fail to align to the human genome. 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.

Number of Mismatches between The 1000 Genomes Project DNA and Mycoplasma

Figure 3 shows that although Bowtie finds matches within one or more Mycoplasma genomes for 75 879 DNA sequences drawn from The 1000 Genomes Project (but does not match them with the human

[1]Some scanners report DNA sequences for both ends of a fragment of DNA. Nonetheless the pair of sequences is considered one "DNA measurement". See also Figure 2.

reference genome) the accuracy of the match varies considerably. Many match a Mycoplasma exactly. These are shown on the left of Figure 3. For others, Bowtie reports up to 78 mismatches. Note the long thin tail to the right in Figure 3. Figure 3 also breaks these data down into pair end and single DNA strands and Solexa coding type (normal v. SOLiD colorspace). Colorspace encoding is described on page 7. 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 9. On average, if affected, colorspace scans contain many more affected DNA measurements than normally coded Solexa scans. See columns 3–4 of Table 1. Overall ten percent of The 1000 Genomes Project scans 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.

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 $\log_{10}$) per base position in each DNA sequence. (For example, a quality of 0.5 ($S/N = 3.16$) means the returned base is more likely than the other 3 combined)[2]. 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. Nonetheless in our large sample of more than 50 billion DNA measurements drawn randomly from The 1000 Genomes Project, there are 1944 measurements with a quality above 0.5 (which match one or more Mycoplasma genomes with ≤ 3 mismatches). They occur in 269 scans, this is 7% of our sample, see last number in Table 2.

[2] Whilst details depend on the individual manufacturer, essentially each base is allocated a different colour. The brightest colour indicates the base and the quality is estimated from how strong it is compared to the other three colours.

Entropy of The 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.

Confirming Bowtie with NCBI BLAST

Rather than trying to run BLAST [10] on several thousand DNA strings, we used 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. BLAST provides strong evidence that these DNA measurements are really from one or more Mycoplasma or similar species.

Conclusions

Here we have analysed DNA sequences directly, rather than gene expression. While the techniques are totally different, there is still considerable scope for sample contamination and sequence comparison, Table 2, 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. [11] and Cibulskis et al. [12], have looked at contamination by other members of the same species. Indeed there have been several reports of unexpected personal, i.e. human, DNA in The 1000 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. Since some DNA sequences have been conserved by evolution, it is possible the contamination is from similar species.

Implications and Future Work

Once Mycoplasma is suspected, it may be that individual scans can be clean up relatively easily as cross-species contamination is said to be easily detected [12, page 2601]. Indeed a number of commercial Mycoplasma detection tools are based on looking for Mycoplasma genes [2]. However both current microbiology laboratory [2] and Bioinformatics [1] typically take the robust approach of removing (deleting) all potentially infected materials. Indeed when The 1000 Genomes Project withdraws nextGen data, it withdraws complete scans. That is, it simply discards information on about a billion DNA bases each time a scan is withdrawn.

Raw data from The 1000 Genomes Project are publicly available and are being increasingly widely and diversely used. Whilst noisy data may be acceptable for use by their original owners, who are aware of their limitations, there is an increasing risk of contaminated data being (ab)-used outside the laboratories which initial created them. Indeed with staff-turnover there may be risks associated with using what becomes historical data where their provenance becomes more cloudy.

Independent numerical studies could be done. The size of our sample suggests (at least for historical data drawn from the same period) they should yield the same results. However, whilst we have established a lower bound for contamination, future studies should be able to calculate it more precisely. For example, by considering redundant scans and clusters it should be possible to isolate the source and perhaps also provide numerical techniques to mitigate the data [12]. Other studies might also look for other effects and thus extract more scientific knowledge from this valuable resource.

Since Mycoplasma are rampant in modern microbiology laboratories [2] it is no surprise to find some in parts of data from The 1000 Genomes Project. We have identified some samples which have a higher than average chance of being contaminated by Mycoplasma. *In silico* studies should be reinforced by checking the source of the data. We urge each member of The 1000 Genomes Project Consortium (as some are apparently doing [12]), particularly those using single ended colorspace scanners (cf. Table 1) to re-check their procedures. Drexler and Uphoff [2] suggest using at least two detection techniques when checking samples for Mycoplasma.

Methods

The master index file, `sequence.index`, which describes all the current 1000 Genomes Project data was down loaded[3]. As of 8 February 2013 there were 47,315 scans available (a further 208 had been withdrawn). They comprised: 39 736 paired-end and 4822 single ended DNA sequence scans plus a further 1611 (paired end) and 938 (single ended) scans which used ABI_SOLID colorspace encoding. 4058 were randomly chosen and down loaded. All the DNA measurements are in fastq format, so they include a quality score per DNA base pair. Each scan contains DNA sequences of the same length. Figure 9 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.

On average: each scan 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 scans 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 [8] 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 10. We used all of the Mycoplasma genomes available from NCBI (30 in total. See Table 4.) 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 that for the two 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.

Notice (Figure 10) 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.

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

Estimating Entropy

In statistical mechanics, entropy is the degree of disorder in a system [13]. In information theory this translates to the degree or randomness or incompressibility of data, particularly in transmission of messages [14]. entropy $= -\sum p \log p$, where p is the probability of a sequence of symbols and we sum over all possible symbols. For replicability, the remainder of this section details how we approximate entropy using actual DNA base counts in finite sequences.

In order to have entropy expressed in bits we use $\log_2$.

A reasonable estimate of the compressibility of variable length DNA sequences can be made by considering all loss-less coding schemes of up to four bases. The most efficient coding scheme gives the most compressible output. For example, a long sequence of adenine (AAAAAAAAAAAAAAAAAAAA...) can be recoded as a shorter sequence of 00000..., where 0 is one of the new 256 codes needed to represent AAAA–TTTT. Since the coding is loss-less, the encoded sequence contains the same information and so it has the same entropy.

We approximate probability p by the actual ratio of each symbol to the number of symbols in the string, $p = i/l$, so entropy $= -\sum_{\text{all symbols}}(i/l)\log_2(i/l)$. Where l is the length of the encoded string and i is the number of each symbol in it. To get the best estimate, we would have to consider all codings. By using the minimum of all 10 possible codings of length up to four DNA bases, we get a reasonable estimate that can deal exactly with not only runs of single bases up to runs of four repeated bases, but gives reasonable estimates with larger repeating sequences. DNA bases which are unknown (i.e. coded as N) are ignored. We use entropy $= \min_{\text{all codings}}(-\sum(i/l)\log_2(i/l))$. Thus the sequence ACGTACGTACGTACGTACGT, which is highly compressible, has an entropy of $-(5/5)\log_2(5/5) = 0$. Whereas a simple count of number of bases would show A C G and T each occur 5 times (are present in equal numbers) and so incorrectly would say the string has maximal entropy $-\sum_{\text{i=A,C,G,T}}(5/20)\log_2(5/20) = 2$. More sophisticated calculations might consider longer potential coding sequences but then the coding tables would be much larger and eventually their information content could no longer be ignored.

Two Base Colorspace Encoding

Some next generation DNA scanners use a technology which instead of reading DNA sequences one base at a time they use multiple fluorescent dyes to read adjacent (overlapping) pairs of bases. Reduced noise is claimed, since as the pairs overlap, each base is read twice. Data are presented as the initial base followed by transitions from one base type to the next in the sequence (hence needing 4 colours). A potential

downside is if an error does occur, the rest of the sequence will be nonsense. Whereas in direct encoding only the erroneous base is effected. It is possible to convert between the two encodings. However because of the different noise characteristics it is usually recommended, as we did, to use tools like Bowtie which can deal with colorspace encoded data directly.

Selecting a high quality Sample to Confirm with NCBI BLAST

We used NCBI's Blast [10] 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 (see page 4) 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.

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, which makes it one of the more frequent genes in today's DNA sequence databanks.

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.

References

[1] Miller CJ, Kassem HS, Pepper SD, Hey Y, Ward TH, Margison GP: **Mycoplasma infection significantly alters microarray gene expression profiles**. *BioTechniques* 2003, **35**(4):812–814, [http://www.biotechniques.com/BiotechniquesJournal/2003/October/].

[2] Drexler HG, Uphoff CC: **Mycoplasma contamination of cell cultures: Incidence, sources, effects, detection, elimination, prevention**. *Cytotechnology* 2002, **39**(2):75–90, [http://dx.doi.org/doi:10.1023/A:1022913015916].

[3] Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Edgar R: **NCBI GEO: mining tens of millions of expression profiles–database and tools update**. *Nucleic Acids Research* 2007, **35**(Database issue):D760–D765, [http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve\&db=pubmed\&dopt=Abstract\&list_uids=17099226].

[4] Aldecoa-Otalora E, Langdon WB, Cunningham P, Arno MJ: **Unexpected presence of mycoplasma probes on human microarrays**. *BioTechniques* 2009, **47**(6):1013–1016, [http://dx.doi.org/doi:10.2144/000113271].

[5] Langdon WB: **Correlation of Microarray Probes give Evidence for Mycoplasma Contamination in Human Studies**. In *GECCO-2013 Workshop: MedGEC Medical Applications of Genetic and Evolutionary Computation*. Edited by Smith SL, Cagnoni S, Patton RM, Amsterdam: ACM 2013:1447–1454, [http://doi.acm.org/10.1145/2464576.2482725].

[6] Langdon WB, Arno M: ***In Silico* Infection of the Human Genome**. In *10th European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics, EvoBIO 2012*, *Volume 7246 of* LNCS. Edited by Giacobini M, Vanneschi L, Bush WS, Malaga, Spain: Springer Verlag 2012:245–249, [http://dx.doi.org/doi:10.1007/978-3-642-29066-4_22].

[7] Durbin, *et al* RM: **A map of human genome variation from population-scale sequencing**. *Nature* 2010, **467**(7319):1061–1073, [http://dx.doi.org/10.1038/nature09534].

[8] Langmead B, Trapnell C, Pop M, Salzberg S: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome**. *Genome Biology* 2009, **10**(3):R25, [http://genomebiology.com/2009/10/3/R25].

[9] Schmieder R, Edwards R: **Fast Identification and Removal of Sequence Contamination from Genomic and Metagenomic Datasets**. *PLoS ONE* 2011, **6**(3), [http://dx.doi.org/10.1371%2Fjournal.pone.0017288].

[10] Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST a new generation of protein database search programs**. *Nucleic Acids Research* 1997, **25**(17):3389–3402, [http://nar.oxfordjournals.org/content/25/17/3389.abstract].

[11] Jun G, Flickinger M, Hetrick KN, Romm JM, Doheny KF, Abecasis GR, Boehnke M, Kang HM: **Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data**. *American Journal of Human Genetics* 2012, **91**(5):839–848, [http://dx.doi.org/10.1016/j.ajhg.2012.09.004].

[12] Cibulskis K, McKenna A, Fennell T, Banks E, DePristo M, Getz G: **ContEst: estimating cross-contamination of human samples in next-generation sequencing data**. *Bioinformatics* 2011, **27**(18):2601–2602, [http://bioinformatics.oxfordjournals.org/content/27/18/2601.abstract].

[13] Pippard AB: *Elements of Classical Thermodynamics*. Cambridge University Press 1957, [http://adsabs.harvard.edu/abs/1957ectf.book.....P].

[14] Shannon CE, Weaver W: *The Mathematical Theory of Communication*. Urbana, Il, USA: The University of Illinois Press 1964, [http://www.press.uillinois.edu/books/catalog/67qhn3ym9780252725463.html].

Figures

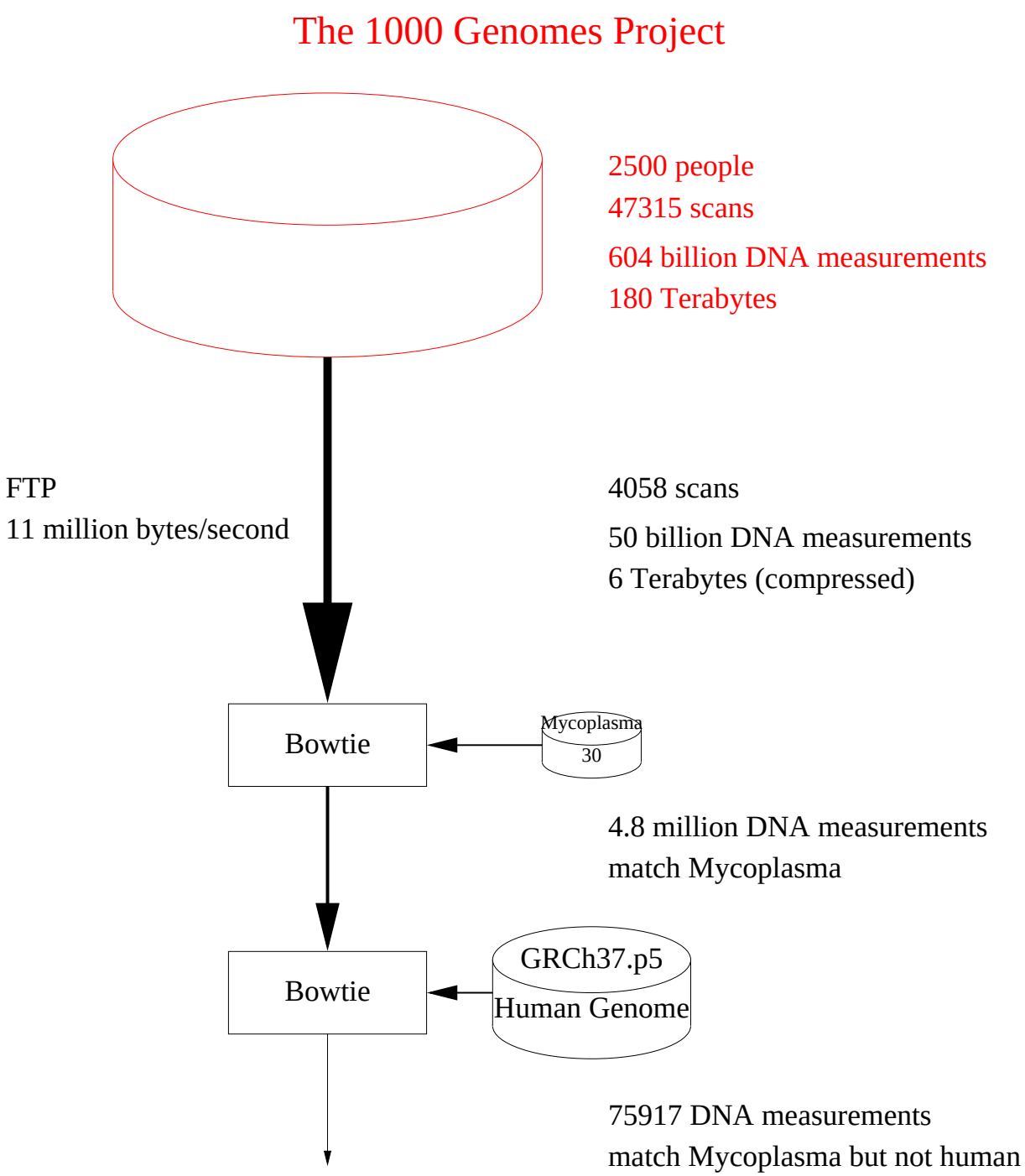

Figure 1: Schematic showing major data flows in Mycoplasma analysis of The Thousand Genome Project (top color). A random sample ($\approx 8\%$) of next generation scan are copied across the Internet to the computer at UCL (black). Bowtie [8] is used to extract individual and paired-end DNA measurements which match one or more of the thirty published Mycoplasma genomes (Table 4). Bowtie is used a second time to exclude DNA measurements which match the reference human genome, leaving 75 879 Mycoplasma DNA measurements from 2055 scans of the 4058 downloaded.

Figure 2: The 1000 Genomes Project uses a variety of next generation DNA sequencing machines (also known as scanners). Some use paired-end DNA strands (schematic above). These scans give the DNA base sequence at both ends (shown as solid black). Only the approximate number of bases between the ends is known. The scanner does not report the sequence of bases between the ends. With paired-end scanners, the two ends together are referred to as a single "DNA measurement". Other scans only contain the sequence of bases at one end of the DNA strand. In these cases there is also one "DNA measurement" per DNA molecule.

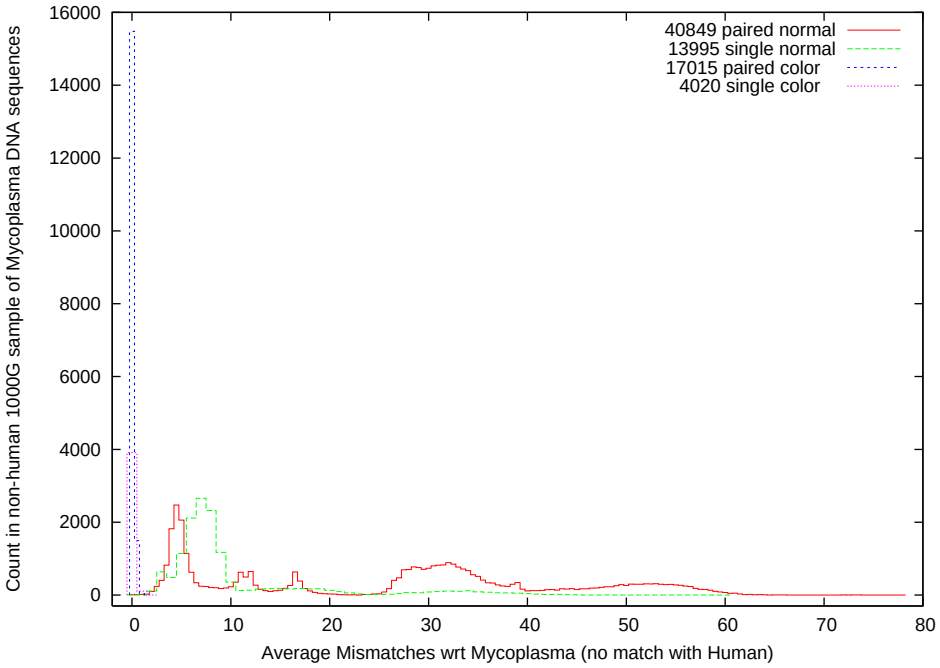

Figure 3: Distribution of mismatches in matches against Mycoplasma genomes for 75 879 DNA sequences from The 1000 Genomes Project which do not match the reference human genome. Of these 40 849 are paired end DNA sequences from Illumina or 454 Life Sciences (not colorspace) next generation scanners, 13 995 are single ended also produced by Illumina or 454 Life Sciences scanners, 17 015 paired end produced by Life Technologies SOLiD colorspace scanners and the remaining 4 030 were also reported by Life Technologies SOLiD colorspace but are single ended DNA sequences.

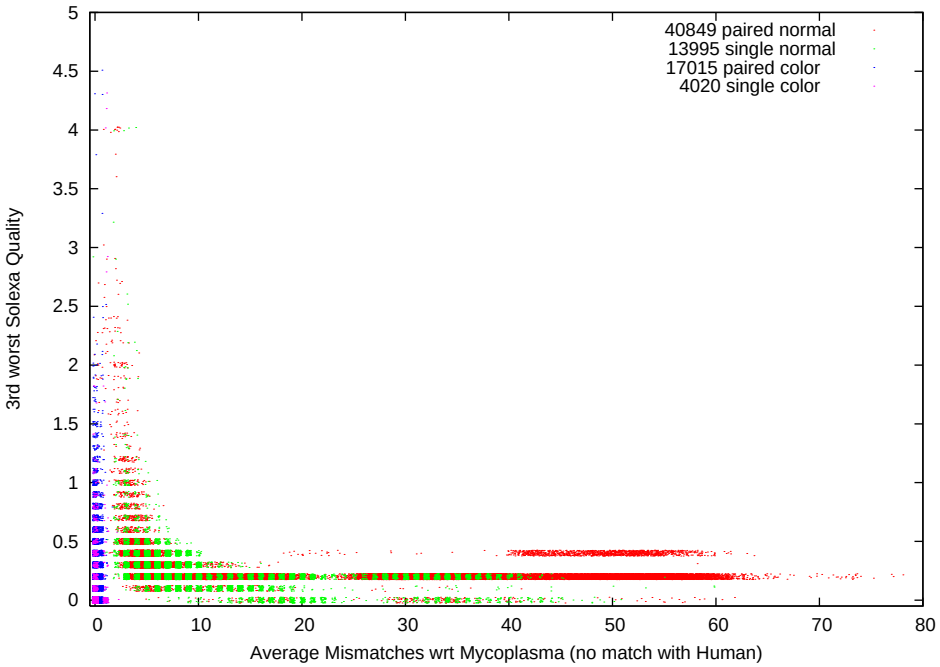

Figure 4: Quality of 75 879 sequences from The 1000 Genomes Project which match one or more Mycoplasma genomes but do not match the reference human genome. Horizontal and vertical noise added to spread data. Most sequences which fail to match GRCh37.p5 but do match one or more species of Mycoplasma are of low quality. Nevertheless an important fraction are of high quality and match Mycoplasma with no or few mismatches. As with Figure 3, data are split in four by type of sample preparation and sequencing machine.

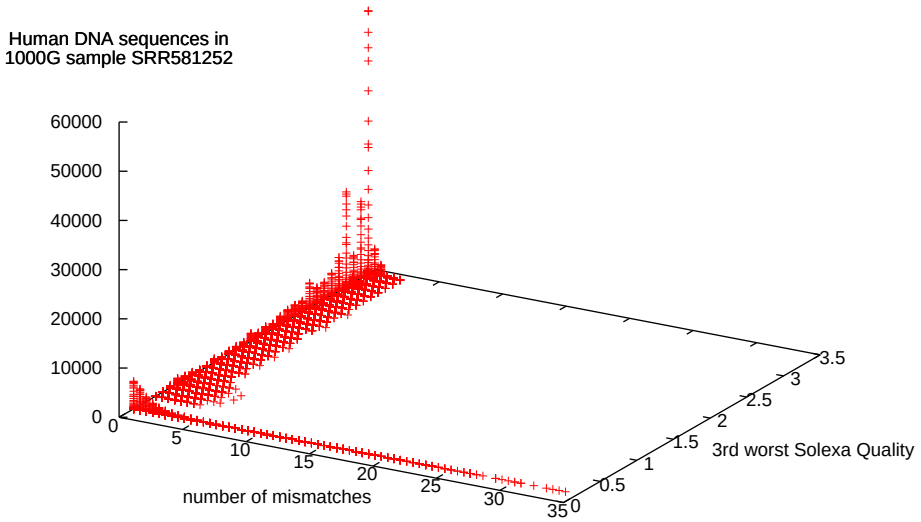

Figure 5: Quality v. accuracy of match (horizontal) for 1 762 302 DNA sequence pairs which match the human reference genome. (From an example 1000 Genomes Project paired-end run.) Showing typically large numbers (> 5) of mismatches are only reported for poor quality data.

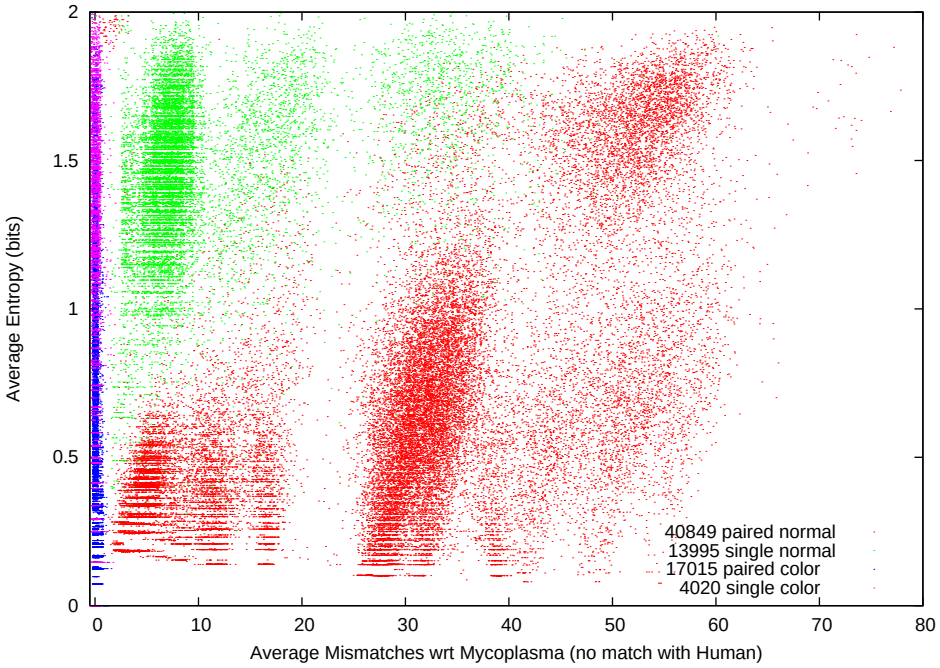

Figure 6: Entropy per DNA base of 75 879 sequences from The 1000 Genomes Project which match one or more Mycoplasma genomes but do not match the reference human genome. (See also Figure 7. Horizontal noise added to spread data.) As with Figure 3, data are split in four by type of sample preparation and sequencing machine.

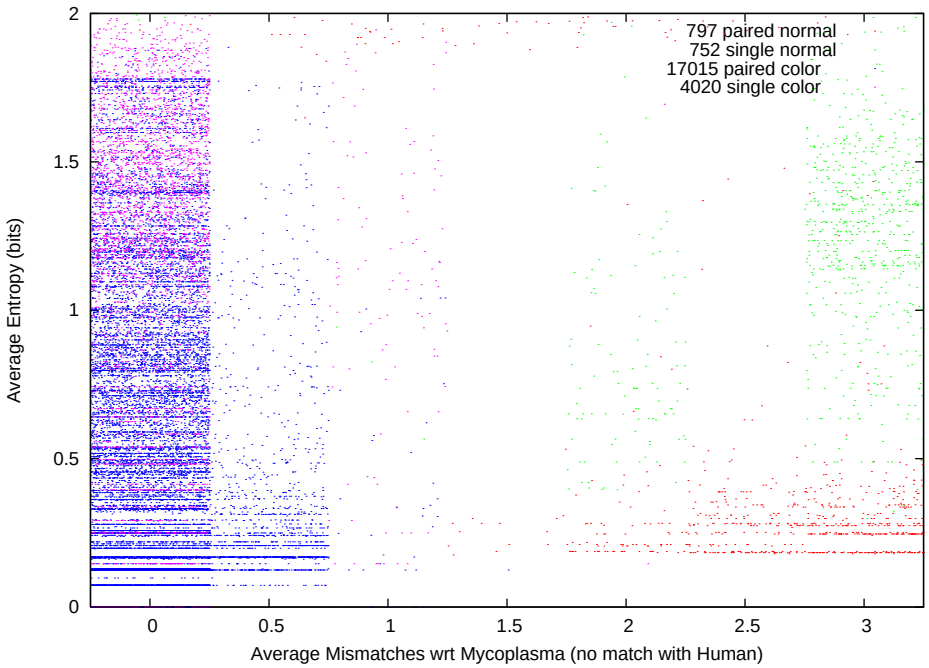

Figure 7: Entropy per DNA base of 22 584 sequences from The 1000 Genomes Project which match one or more Mycoplasma genomes with 3 or fewer mismatches but do not match the reference human genome. (Detail of Figure 6. Horizontal noise added to spread data.) As with Figure 3, data are split in four by type of sample preparation and sequencing machine.

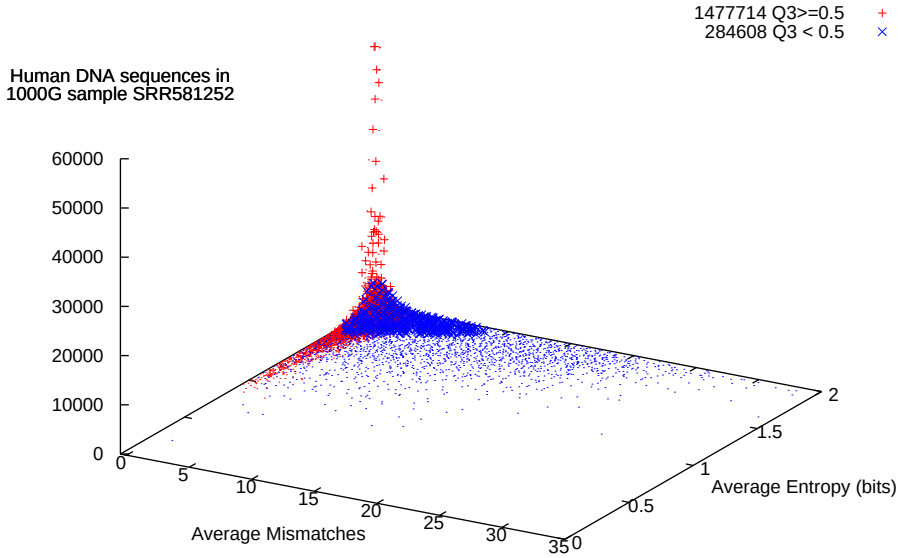

Figure 8: Entropy v. number of mismatches for 1 762 302 DNA pairs which match the human reference genome. (From the same example 1000 Genomes Project paired-end run as in Figure 5.) Most DNA measurements which match GRCh37.p5 are not repetitive (i.e. have high entropy). Also low quality (×) measurements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

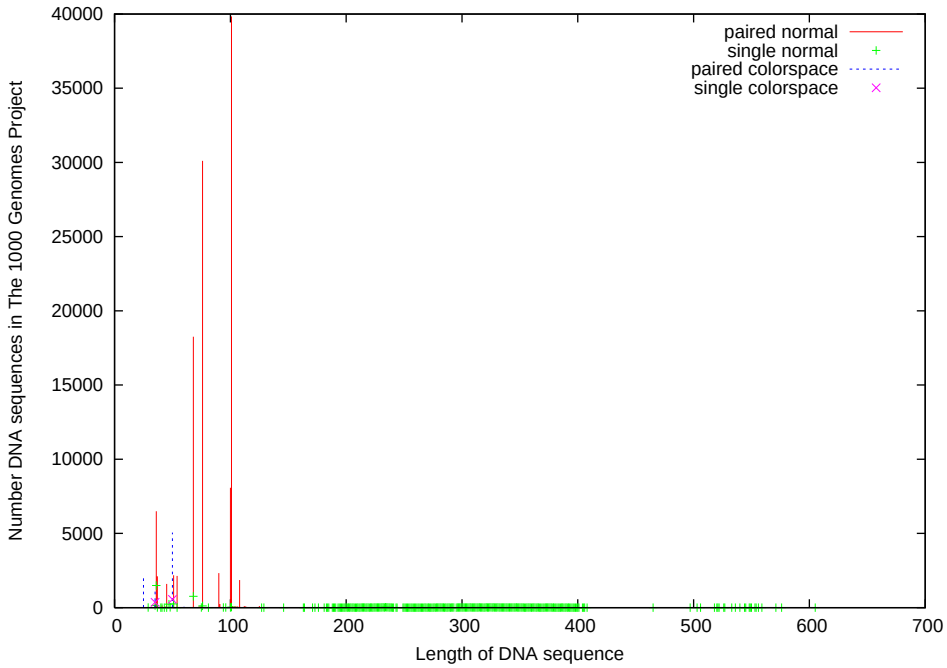

Figure 9: Lengths of DNA sequences for The 1000 Genomes Project. Mostly measurements have two paired ends. The mode is for each end to have 101 DNA base pairs. Again, as with Figure 3, data are split by type of sample preparation and sequencing machine.

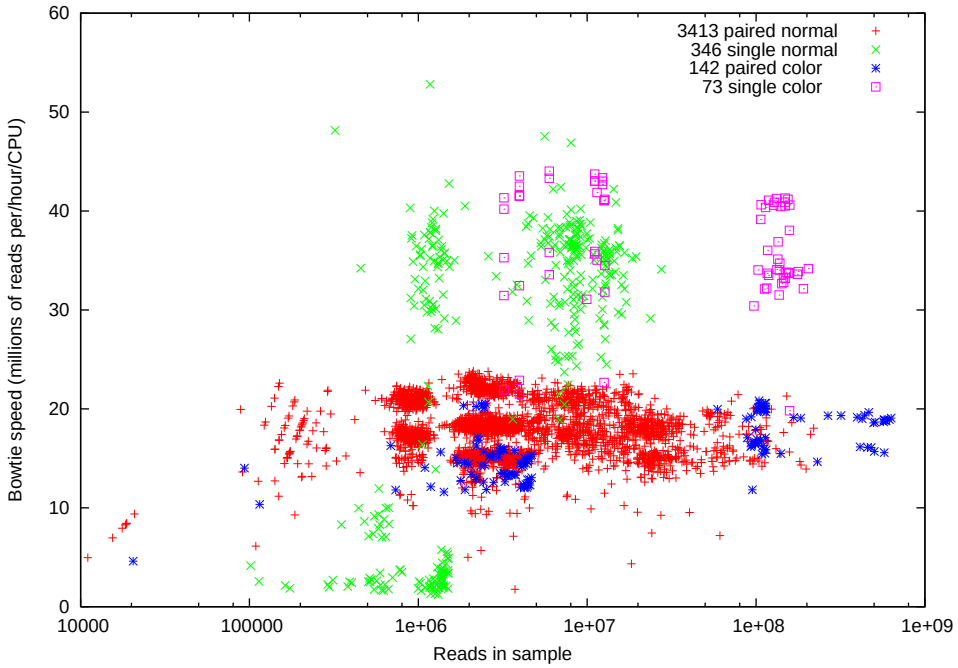

Figure 10: Speed of Bowtie mapping short nextGen DNA sequences from The 1000 Genomes Project against 30 Mycoplasma genomes (Table 4). As expected, Bowtie is typically about twice as fast on single ended ($\times$ and $\square$) compared to paired end DNA sequences ($+$ and $*$). For the colorspace DNA sequences ($*$ and $\square$) Bowtie uses a colorspace version of its EBWT indexes. Bowtie run on a 32 GB 8 core 3 GHz server. Note log horizontal scale.

Tables

Table 1: Approximately 8% of The 1000 Genomes Project was selected at random and downloaded. Of the 53 billion DNA measurements down loaded, 22 584 (bottom column 3) match Mycoplasma (with on average three or fewer mismatches) but do not match at all the reference human genome GRCh37.p5. Typically samples infected with Mycoplasma are immediately destroyed. Column 4 gives the number of scans with at least one affected DNA measurement. The final column gives this as a percentage of scans of the same type.

Type		Mycoplasma sequences	Affected scans	Scans downloaded	Fraction of scans
pair	ordinary	797	106	3454	3%
pair	colorspace	17 015	111	145	77%
single	ordinary	752	108	384	28%
single	colorspace	4 020	72	75	96%
	Totals	22 584	397	4058	10%

Table 2: In our random download of The 1000 Genomes Project, 1944 high quality DNA measurements (i.e. no more than three bases with quality worse than 0.5) match Mycoplasma (with on average three or fewer mismatches) but do not match at all the reference human genome (bottom column 3). As with Table 1, column 4 gives the number of scans with at least one affected DNA measurement. The right most column gives the percentage of affected scans by type.

Type		High quality Mycoplasma sequences	Affected scans	Scans downloaded	Fraction of scans
pair	ordinary	542	87	3454	3%
pair	colorspace	1042	63	145	43%
single	ordinary	234	78	384	20%
single	colorspace	126	41	75	55%
	Totals	1944	269	4058	7%

Table 3: High quality, non-repetitive DNA measurements from The 1000 Genomes Project 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). E= chance of random match [10]. Column 5 gives an example gene (see discussion on page 8).

2.0	2.9	ERR009050. 2605525	GCCGTAACTATAACGGTCCTAAGGTAGCGAAATTCCTTG TC E=7 10^{-13}	S16 23S ribosomal RNA
2.0	2.3	ERR002459. 4464466	ACGGTTTTCAAGACCGTTCCCTTCAGCCAGACTTGG E=5 10^{-10} CCTGACGGTTTTCAAGACCGTTCCCTTCAGCCAGAC E=5 10^{-10}	transfer RNA-Ser
1.8	2.1	ERR013159. 14600701	CGCTTTCATTGTTCCGCCAGTAGCTAAAACATCATCAATA ATTGCTACTTTTTGGCCTTTTTTCAACATATTAGTTTGGA TTTCTAGAGTTGATTTACCATATTCTAA E=3 10^{-52} TTTTTGGCCTTTTTTCAACATATTAGTTTGGATTTCTAGA GTTGATTTACCATATTCTAAATCATACTCAAAACTAATAA CGTCTCCTGGTAATTTTTTAGGTTTTCT E=3 10^{-52}	
1.8	2.0	ERR013159. 12593030	GAGCTTGTTTTTCGTATTTTTCAATTTCTATTTCGTCATT GATTTGTCAATTTGGTAAATTTGTGTTTTCGCTATCAGGT TTGGTTAGTTTAAAATAACCATCAAAAG E=2 10^{-10} AGGTTTGGTTAGTTTAAAATAACCATCAAAAGTAATTATT GAACCAGAAAGATAAAATTTGTGTTCTTGATTTAAAAATT CATAACGTGTAATTTGTCTTTCAGGAAC E=3 10^{-52}	
1.7	2.2	ERR013159. 18901091	GGTCAAGTTTACAACAAAATGTTTGCACTTCAAAAAGAAC TAGAAGAACTAGAAGAAAATAAAGAAGAAAATACTTTAA TCAAAGAAGTAGTGAACCAAGAAGATATT E=3 10^{-6} AAAAGAACTAGAAGAACTAGAAGAAAATAAAGAAGAAAA TACTTTAATCAAAGAAGTAGTGAACCAAGAAGATATTGCA AATATTGTTTCTAAATGAACAAAAATTCC E=3 10^{-9}	
2.0	1.9	ERR013159. 7037432	TCTAGAGATACTGCCTGGGTAACCAGGAGGAAGGTGGGG ACGACGTCAAATCATCATGCCTCTTACGAGTGGGGCAACA CACGTGCTACAATGGTCGGTACAAAGAGA E=3 10^{-52} AGTGGGGCAACACACGTGCTACAATGGTCGGTACAAAGA GAAGCAATATGGTGACATGGAGCAAATCTCAAAAAACCG ATCTCAGTTCGGATTGAAGTCTGCAACTCG E=3 10^{-52}	16S ribosomal RNA
1.9	1.9	ERR022473. 14544768	TGCTTTTTTACCTCATGGAGTAAGTGGTGCTTTACGTCCA ATTGGTTGTTTACCTTCACCACCACCATGTGGGTGATCAT TTGGGTTCATTACAGAACCTCTAACTGT E=3 10^{-52} GGTGATCATTTGGGTTCATTACAGAACCTCTAACTGTTGG ACGAATACCTAAATGACGATTACGTCCTGCTTTTCCAATG TTAACTAGGTTATGTTCTTCATTTCCTA E=3 10^{-52}	ribosomal protein cluster

Additional Files

Mycoplasma Genomes Used

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

Table 4: Thirty species of Mycoplasma whose Genomes were used

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