A thousand genomes and then some

The 1000 Genomes Project was established as a way of cataloging human genetic variation, but it seems to have inadvertently cataloged something else - Mycoplasma genomes - this is according to a study published in the open access journal *BioData Mining*.

Raw data from the 1000 Genomes Project were analyzed by the single author, William Langdon, a computer scientist from University College of London. He downloaded 50 billion DNA measurements from The 1000 Genomes Project and found that some of the data did not match human genomes. When further analysis was carried out, including BLAST searches, it was confirmed that these genomes were Mycoplasma and occur in at least 7% of the samples.

Mycoplasma contamination is a common problem in labs usually caused by infection of the medium that cells are grown in. When contamination is detected in data it is quite easy to clean up subsequent scans taking this into account. Stephan Beck, professor of medical genomics at University College London, says: "When scientists download these raw data from The 1000 Genomes website, or any similar project, they should be aware of the caveat that this data is exactly that - raw. It is not surprising that contamination was found, but this should act as a warning to the community that they need to more vigilant and filtering out this contamination."

The 1000 Genomes Project was launched in 2008 as way of cataloging the different variants that results in genetic diversity in humans. It hopes to achieve its goal by sequencing the genomes of more than 1000 people. These data are then uploaded to the Internet where other researchers are free to use them. The data are used in one of two ways, firstly, scientists check how common a variation is in a certain condition across a population instead of comparing to one human reference genome. Secondly, it is possible to check the variation within different ethnicities to study human population history, as well as what impact ethnic variation may have upon disease.

This unwanted appearance of Mycoplasma genomes could also be seen as an opportunity. Once these genomes are removed they could then be analyzed by specialists in the field of Mycoplasma research. This could then give a better understanding of how Mycoplasma acts in growth medium.

William Langdon says: "As scientists use publicly available datasets rather than collecting their own samples, there is a risk of people using data and taking it as gospel. Mycoplasma contamination is a common problem, but it's just a case of catching it and annotating data."