EMBO reports

# Here, there, and everywhere

*From PCRs to next-generation sequencing technologies and sequence databases, DNA contaminants creep in from the most unlikely places*

Karl Gruber

The main hurdle for genome sequencing projects these days is no longer the effort and cost of generating sequence data—which has become exponentially cheaper—but the capacity to analyse huge amounts of data and make sense of it. This endeavour is made harder by another problem that has begun to emerge over the past years: DNA contamination. Contamination impacts both sequence data generation, when DNA from other species finds its way into samples or equipment, and analysis, when DNA sequence data from contamination finds its way into databases. While researchers are usually aware of the potential for the first problem—DNA is an ubiquitous and hardy molecule that can persist anywhere in the laboratory and appropriate measures are usually taken to prevent contamination—database contamination has not yet triggered serious concerns, let alone measures to deal with it. The great challenge is not just to avoid DNA contamination in the first place—during PCR amplification and sequencing—but also to identify tainted sequence data in important data repositories.

> *"The great challenge is not just to avoid DNA contamination in the first place [. . .] but also to identify tainted sequence data in important data repositories."*

Many "wet" laboratories are aware of the problem of potential contamination and have adopted precautionary measures to avoid it, in addition to the normal negative and positive controls. DNA extractions and PCR setup are done at a designated location, whereas PCR amplification and analysis of the amplicon are performed in a separate room. In fact, people are usually not allowed to bring anything from the PCR laboratory back to the DNA extraction or PCR setup room. Despite such efforts, contamination still shows up, especially in laboratories where many researchers work on different species and use universal primers. For example, if someone works with high-quality tissue that yields highly concentrated DNA, while others work with samples that have much lower amounts of genomic DNA, the high-yield DNA can often get into the low-yield samples just because there is so much of it. "In my own experience with 16S rRNA and metagenomic studies of bacterial populations, background levels of contaminating taxa are often observed. These taxa are mostly soil and water bacteria, originating from DNA extraction kits, laboratory reagents and ultrapure water," commented Susannah Salter from the Wellcome Trust Sanger Institute in the UK, who works on pathogen genomics. Bioinformatician Tom Slezak, from the Lawrence Livermore National Laboratory, in California, USA, noted that even sequencing machines carry their own risks of contamination, as DNA left behind from previous sequencing experiments may persist, "despite however many stringent bleach rinses are employed above and beyond what the manufacturer suggests," he said. But the biggest problem in Slezak's view is so-called 'naked DNA': "Once a sample is present in a laboratory, there will be some amount of DNA that wafts around and becomes part of the laboratory environment. Such DNA may eventually contaminate subsequent samples," he explained.

> *". . . even sequencing machines carry their own risks of contamination, as DNA left behind from previous sequencing experiments may persist. . ."*

Sterilization of working reagents and reusable plastics is a standard practice in most laboratories, but sterilization will not necessarily destroy DNA, Salter explained. "For example, autoclaving plastic or glassware will kill the contaminating bacterial cells, but the DNA released from those lysed cells will still be present. The autoclave itself may become the source of DNA contamination for the lab," she said. "Bacterial DNA may also be present in buffers, reagents and products brought in from other sources, so copious negative controls are a good idea to spot this when it arises." Many laboratories therefore use bleach or UV irradiation to destroy DNA in critical areas.

The most extreme measures are taken in laboratories that try to obtain sequence data from old human or animal remains. Contamination is basically an unavoidable problem, according to Matthias Meyer, group leader at the Max Planck Institute for Evolutionary Anthropology in Leipzig, who successfully recovered DNA sequences from the bone of an archaic human who lived 400,000 years ago. "Not only are most of the bones already contaminated when we get them, we also cannot avoid the introduction of some additional contamination during lab work," Meyer said.

Freelance journalist in Perth, WA, Australia. E-mail: karlgruber@gmail.com
**DOI** 10.15252/embr.201540822 | Published online 6 July 2015


© 2015 The Author

How do you sequence the genome from an ancient bone sample that contains only minute amounts of DNA from the species itself? Very carefully. Before entering the laboratory, Viviane Slon, a PhD student in Meyer's group, puts on a disposable whole body suit, a hair net, a face mask, a face shield and two pairs of gloves: "One pair of gloves is the 'baseline', which never comes off, while the second one is disposable—we change gloves constantly in the clean room," she explained. "The idea is to not have any skin exposed while working with the ancient sample." Upon entering the clean room, where she processes ancient samples to extract DNA, Slon turns off the UV light that is constantly turned on when the room is not in use. The clean room itself is actually three separate rooms, each devoted to a specific step of sample processing. In the first room, Slon drills into bone or teeth in a dedicated fume hood, using disposable sterile dentistry utensils. Her first task is to remove the surface of each bone, as the people who originally found them usually handled them with their bare hands, resulting in contamination.

Slon said that paranoia pays off when working under the fume hood. Tubes, reagents and pipettes are arranged in such a way that her arm never pass over an opened tube. "In the hood, never have more than one tube opened at a time, and never pass

© 2015 The Author

over with your arm an opened tube; to avoid shedding particles from your sleeves into the tube, for example," she explained. "The inside of the cap has DNA on it, which we don't want to transmit to the next tube we open. When in doubt—change gloves," she added. This applies to reagents as well. "…as with the gloves, when in doubt—change it. If I ever find myself wondering whether a reagent is clean […] it would immediately be thrown away and replaced." On a given work day, Slon can end up using two full boxes of gloves. "Every time you want to put your hands under the hood—you first change gloves. If you have any doubts whether you touched something that could be contaminated (that includes the bone itself), you change gloves," she said. The reason for these extreme precautions goes beyond contamination. "For some of the samples we process, we only get one shot (some are very tiny fragments of bone/tooth). So I am also very careful to not make any errors while working," Slon said.

*"There is growing concern about how much contamination there is in the billions of DNA bases that are currently published and uploaded to databases"*

Throughout the whole process, from DNA extraction, to PCR and sequencing, positive and negative controls are used to ensure that the correct DNA is sequenced. Yet, despite the preventive measures, "contamination often adds up to the equivalent of about 1% of the DNA in a human cell, which makes it hard to generate clean sequence data from samples containing less than one or two copies of the nuclear genome," Meyer said.

Contamination has also found its way into published papers. Meyer re-analysed a published data set claiming to have recovered ancient DNA from a 12,000- to 13,000-year-old human tooth found in an underwater cave in the Yucatan peninsula, Mexico [1]. The finding was amazing, as so far no one had been able to recover ancient DNA from tropical samples, where DNA quickly degrades. But may be too good to be true, as Meyer's analysis concluded that rather than an amazing discovery, the findings could also be explained by two instances of contamination: of the bone or sampling equipment with Native American DNA, and cross contamination with library molecules from a different ancient sample [2]. New techniques that work at a nano-scale may help to better address contamination. "One promising strategy to overcome this problem is the reduction of reaction volumes to nanolitre scale, as achieved by microfluidic systems in the context of single-cell sequencing, which greatly reduces the contribution of contaminant DNA from reagents," Meyer explained.

Sequencing a whole human genome can cost as little as a thousand dollars these days. This dramatic reduction in prices has been met with an explosion of projects that are generating and analysing genome-wide sequence information. The public genome database GenBank now harbours more than 27,000 prokaryotic and 1,600 eukaryotic genomes. A simple search in PubMed for the word "genome" will return more than 40,000 publications in 2014 alone. Filter these results to show only clinical trials, and PubMed shows at least 326 clinical trials that involve genome information in 2014.

Next-generation sequencing technologies have also achieved such a high sensitivity that the tiniest amounts of DNA can be detected and sequenced. This is both good and bad. Good if it targets the right DNA. Bad if tiny amounts of contaminating DNA are sequenced. The problem escalates when contaminants go undetected in negative controls, as erroneous sequence data can find its way into databases and beyond. A good example of such un-detectable contamination was reported recently when researchers discovered that a new DNA virus, first thought to be present in seronegative hepatitis patients [3], was actually a contaminant in the DNA extraction columns.

There is growing concern about how much contamination there is in the billions of DNA bases that are currently published and uploaded to databases. This question is particularly important in regard to medical research, where sequence data may contribute to treatment decisions, drug design or diagnosis. Indeed, post-sequencing analyses have shown the presence of contaminating DNA from other humans and microorganisms in sequenced DNA samples from patients. In the past few years, various reports have highlighted cases of DNA contamination in published genome data [4–6], suggesting that DNA contamination may be more widespread than previously thought.

*"Mycoplasma genes not only contaminated software but have been burnt into hardware"*

One major example of contamination in published data sets was recently found in the "1000 Genomes Project". William Langdon, from University College London, UK, re-analysed the published data sets and discovered that 7% of the samples published contained genes from *Mycoplasma* or related species [7]. Langdon highlighted that in this instance, contamination went beyond a DNA database: "…not only had contaminated *Mycoplasma* genes been uploaded into the reference human genome, but they had been copied worldwide and become incorporated into an affymetrix genechip probeset. *Mycoplasma* genes not only contaminated software but have been burnt into hardware," said Langdon. If a high-profile endeavour such as "The 1000 Genomes Project" is affected, could other genome projects be in the same boat? "As people get more remote from the data collection point they may know less about how it was collected and are more likely to take it on trust," Langdon said. Part of the problem is that the erroneous sequences persist. "I think it's fair to say that the initial response of NCBI was something like 'We are just holding the data for the community. It's up to the guys who upload noise to deal with it.'" However, he believes this will change as curators become more proactive and check the integrity of data sets.

Richard Lusk, at the University of Michigan, Ann Arbor, USA, has also looked into DNA contamination among genomic datasets [8]. He focused his attention on a 2013 study, led by Sándor Spisák from the Hungarian Academy of Sciences in Budapest, Hungary, claiming that genes from foodstuff survived digestion and entered our bloodstream [2]. "I was sceptical of that manuscript, since it seemed to be making an extraordinary claim, and that's what sparked my interest in doing a couple of computational experiments myself," Lusk explained.


© 2015 The Author

He knew from colleagues working with ancient DNA that contamination can be extremely hard to account for and that samples with trace amounts of DNA, like the ones from Spisák's study, are a fertile ground for contaminants. He performed his own bioinformatic analyses to determine whether food genes could pass into our bloodstream, or whether contamination was a more likely explanation. "I had to keep in mind that I couldn't actually pinpoint the source of anything I found there—DNA doesn't come with a tracking code to show you where it's been—but I could look for species that are much more commonly found on the skin than in the gut," he said.

> *"As people get more remote from the data collection point they may know less about how it was collected and are more likely to take it on trust"*

Lusk started downloading sequence data from samples with no plausible connection with food, such as individually washed cells. "Food DNA couldn't get to those from the digestive tract, and so if I found food DNAs in those samples, I knew that they had to originate from contamination," he explained. For each DNA sequence purportedly coming from a food source, Lusk searched for similar sequences in NCBI genome database to get an idea of what species could have potentially contaminated the samples. He was able to identify sequences from a wide range of species, all of which were similar to those from common foods we eat. "It turns out that contaminants can come from a very, very wide range of species," Lusk said. Among those he identified were DNA sequences from two microbes that commonly inhabit human skin, one associated with dandruff and another with acne. "DNA is a reasonably durable molecule that's produced in quantity by everything that's lived, and high-throughput sequencing is a very sensitive technology. I was essentially trying to run the negative controls that Spisák *et al* didn't have," Lusk said.

A third example of contamination deals with an intriguing genome-wide association study that suggested prostate cancer was caused by a fungus [9]. Martin Laurence, from Shipshaw Labs, Montreal, Canada, was keen on discovering the identity of this potentially cancer-causing fungus and sequenced a part of the microbe's ribosomal DNA [10]. "I amplified and sequenced ribosomal DNA in a clinical specimen which matched nearly perfectly with Ichthyosporea sp. ex *Tenebrio molitor* (GenBank entry JN699061.1), a sexually transmissible fungus recently found in mealworms. I thought this could be it," he said.

However, after running the sequence through a simple database search at the NCBI website, everything changed. "Unfortunately, I ran the Ichthyosporea sp. ex *Tenebrio molitor* sequence through NCBI BLAST, and found it matched almost exactly to *Malassezia globosa*, a fungus found on the human skin and also a very common laboratory contaminant." According to Laurence, when the JN699061.1 sequence was first uploaded into GenBank's database in 2012, neither the authors nor the GenBank curators ran the sequence against their database. When Laurence did, the results were enlightening. "[I]t matched almost perfectly with AAYY01000016.1, the *Malassezia globosa* genome published in 2007. And *Malassezia globosa* is known to be a common contaminant, like *Propionibacterium acnes*, due to its presence on the human skin," he explained.

Earlier this year, Laurence reported other examples. "In my *PLoS One* paper [10], we report that bacterial genus *Bradyrhizobium* sequences can be found in GenBank entries of various eukaryotes. It can also be found in many 1000 Genome Project sequencing runs, strongly suggesting that it is a laboratory contaminant," he said. "I'm not sure how to go about fixing them (this is a rather large undertaking), but researchers doing microbiome analyses such as myself would really appreciate having reliable taxonomic information in GenBank."

The problem is not restricted to human sequence data. Many microbial genome sequences contain human contamination, Slezak explained. "This is particularly true for some microbial draft genomes that are submitted in bulk without adequate screening," he said. Likewise, microbial genomes may contain contamination from other microbial DNA, as Slezak found recently in a microbial draft genome he analysed. "Such sloppiness on the part of DNA submitters is inexcusable, but unfortunately all too common," he added.

Contamination is likely to remain a regular peril for all brands of DNA-based research. Anyone dealing with such research, from DNA sequencing to data analysis and mining, therefore always has to consider the possibility of contamination. Unfortunately, there is currently both a lack of awareness among many scientists and a lack of efforts to recognize and weed out contaminating sequence data.

## References


1. Chatters JC, Kennett DJ, Asmerom Y, Kemp BM, Polyak V, Blank AN, Beddows PA, Reinhardt E, Arroyo-Cabrales J, Bolnick DA *et al* (2014) Late pleistocene human skeleton and mtDNA link paleoamericans and modern Native Americans. *Science* 344: 750 – 754
2. Prüfer K, Meyer M (2015) Comment on "Late Pleistocene human skeleton and mtDNA link Paleoamericans and modern Native Americans. *Science* 347: 835
3. Naccache SN, Greninger AL, Lee D, Coffey LL, Phan T, Rein-Weston A, Aronsohn A, Hackett J Jr, Delwart EL, Chiu CY (2013) The perils of pathogen discovery: origin of a novel parvovirus-like hybrid genome traced to nucleic acid extraction spin columns. *J Virol* 87: 11966 – 11977
4. Longo MS, O'Neill MJ, O'Neill RJ (2011) Abundant human DNA contamination identified in non-primate genome databases. *PLoS ONE* 6: e16410
5. Witt N, Rodger G, Vandesompele J, Benes V, Zumla A, Rook GA, Huggett JF (2009) An assessment of air as a source of DNA contamination encountered when performing PCR. *J Biomol Tech* 20: 236
6. Merchant S, Wood DE, Salzberg SL (2014) Unexpected cross-species contamination in genome sequencing projects. *Peer J* 2: e675
7. Langdon WB (2014) Mycoplasma contamination in the 1000 Genomes Project. *BioData Min* 7: 3
8. Lusk RW (2014) Diverse and widespread contamination evident in the unmapped depths of high throughput sequencing data. *PLoS One* 9: e110808
9. Sutcliffe S, De Marzo AM, Sfanos KS, Laurence M (2014) MSMB variation and prostate cancer risk: clues towards a possible fungal etiology. *Prostate* 74: 569 – 578
10. Laurence M, Hatzis C, Brash DE (2014) Common contaminants in next-generation sequencing that hinder discovery of low-abundance microbes. *PLoS ONE* 9: e97876


© 2015 The Author