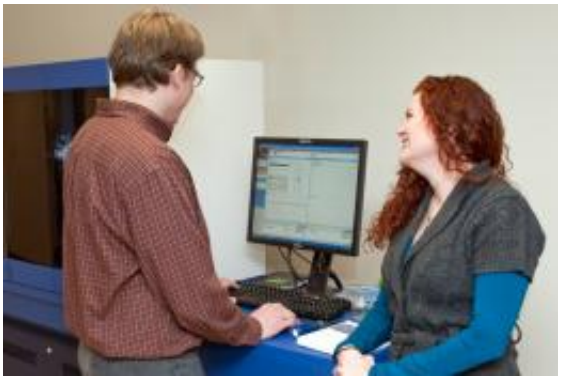
Contamination found in nearly a quarter of genome databases

February 18 2011, By Christine Buckley



Mark Longo, a Graduate student in molecular and cell biology, and associate professor Rachel O'Neill. Photo by Dan Buttrey

(PhysOrg.com) -- UConn scientists say the results could complicate disease identification in humans.

A new genomics study by molecular biologists at the University of Connecticut has shown that at least 22 percent of non-human genome databases are contaminated with human DNA. Their results imply that this level of contamination could also exist in records of the human genome, which could produce major problems in identifying human diseases.

Associate professor Rachel O'Neill, graduate student Mark Longo, and associate professor Michael O'Neill of the molecular and cell biology

department in the College of Liberal Arts and Sciences published their findings today in an online edition of the journal *PLOS One*.

Longo says that he had originally been scanning the genome of zebrafish and comparing it with the human genome to find what are called ultraconserved regions, or bits of DNA that are so ancient they are similar among species that are distantly related, like humans and fish.

But, to Longo's surprise, he found a region of DNA that was identical to one in humans and couldn't be a part of the fish genome. That's when he knew that the fish genome database he was using was contaminated.

"Contamination in these databases could be from people's skin or hair, or it could be DNA from other sequence libraries kept in the same facility," says Longo. "We knew we needed to quantify this to see how many of the databases contained human contamination."

The researchers gathered sequences from all the major global DNA repositories, including the archives at the National Center for Biotechnology Information, the University of California Santa Cruz, the Joint Genome Databases, and the Ensembl genome browser. Any sequencing project funded by federal funds is required to be deposited in one of these archives.

Using a section of DNA that is specific to primates and abundant in the human genome, the researchers identified 454 non-primate genomes out of the 2,027 they sampled as contaminated with human DNA.

Rachel O'Neill says this result led them to reason that if these non-human genome databases were contaminated with human DNA, then it's just as likely that many human databases would be contaminated as well. But, she says, the catch is that it's virtually impossible to identify a foreign bit of human DNA in a human genome database.

“In sequencing, you have to put all the pieces of the genome together like a big jigsaw puzzle. The pieces that don’t fit stand out,” Longo says. “But if you’re working on a human puzzle, it’s like working on a three-billion piece puzzle, and it’s all black.

“It’s virtually impossible to find human contamination in human genome databases,” she adds, because they simply don’t stand out as anything unusual in a human genome. This, she says, could lead to some terrible mistakes.

A portion of the National Center for Biotechnology Information includes a Cancer Genome Atlas: a library documenting mutations that occur in cancer cells. O’Neill says there’s no room for error in these databases.

“It would be very upsetting to be told you have a mutation for breast cancer, when in fact you don’t, and it was just a contamination from another sample,” she says.

O’Neill emphasizes that scientists need to exercise extreme caution when performing their sequencing, and that they should validate results through tests in their own laboratories before submitting them to databases. Longo points out that the UConn researchers found contaminations in some sequences that they had produced in their own laboratories, which they then discarded. O’Neill says these practices should be the norm.

“We’re compounding this problem in our rush to move forward with genomics,” she says. “Millions of dollars are invested each year in these sequence databases, but we’re plowing ahead with less caution than we should. The result is that we might have a harder time recognizing the etiology of something like cancer.”

Longo notes that in his analysis, there was one type of DNA database

that showed no contamination at all: that of influenza. Because viruses are so dangerous, great care is taken in their preparation, he says – much more than is usually taken with a commonplace and harmless genome. This kind of caution should be extended to all sequencing, says O'Neill.

"The sequencing world has moved in leaps and bounds," she says. "It's time for validation to catch up."

Provided by University of Connecticut