**Peer Review # 2**

PhD DNA Genome Sequencing Expert Comments:

Email to Dr. Greer

**April 24, 2018**

Steve,

…. The authors state that the specimen was human and then reference their supplemental note as the source of the evidence for that claim. Their supplemental note is attached to this email for your reference.

This is the only part of the work that actually matters and they stuck it in the supplemental material. There are no references or other supporting figures, tables, or substantive quantitative information for the claims they make in the supplementary note. They follow three lines of inquiry to establish that Atacama is human:

1. High resolution photography

2. X-ray & Computed Tomography

3. Whole genome sequencing

For points 1 & 2, I do not understand how either of these two things could contribute to the case the organism is human. They took fancy pictures of it and described its structure. All the primates have orthologous structures and somehow the scientists are able to classify them into different species. In fact, what the supplemental note actually says is that the organism has non-human features, some with no known human precedent.

Then, we get to point 3, whole genome sequencing (WGS), my bread and butter.

A. They state in the supplemental and repeatedly in the main article that the genomic DNA was of “high quality”. OK. Fine. Where is the evidence to support that? We are just left to take it on faith that the DNA is good? No A260/280. No PicoGreen. No photographs of electrophoresis. No NanoDrop plots. No Sanger dideoxy sequencing runs. All their claims rest on this foundation, Steve. If the quality of DNA was high, then the read quality is high… in other words, the raw data output from the sequencing should also be high. If that is the case and only 91% of the reads mapped (a number far higher than the 70-80% that Gary told me over the phone), then how is that evidence of human? It’s like they’ve placed a blue banana on a table, pointed at it and said, “Yellow banana.” Hey, guess what, the emperor has no clothes. The provide no evidence to support the conclusion that the DNA quality was good.

B. The authors used three Illumina (only) sequencing platforms to generate raw sequence reads. This section is so badly written that I wonder if the PI’s A) even looked at what their grad student wrote or B) looked at it very closely and intentionally made it incomplete, obscure and otherwise difficult to follow. Which of these three platforms produced the 560 million paired end reads they used to determine the organism was not human? We’re left (again) to take it on faith that the reads do pass filter. Fine. Even if we accept that, the numbers don’t add up correctly. (3,300,000,000 base pairs in the human genome)\*(19.6X coverage [reads/base]) = 64,680,000,000 reads. Hmmmmm. 0.56 billion versus 64.7 billion. What? OK, let’s just say that somehow I am doing my math wrong or my understanding of this subject is weak (it is strong) and that they do map 509,000,000 reads to a HUMAN genome reference. This is a key point: 91% of the reads map to a human genome reference and that does NOT mean that 91% of the human genome is covered by reads. If 91% of the human reference genome was covered by reads at 17.7X depth as they incorrectly state, there would be 58,410,000,000 reads, not 509,000,000 as they state. It’s cool, they’re only off by 57,901,000,000. Good enough for government work. So, let’s just say that they are correctly reporting that 509,000,000 reads correctly map to build 37 and that the mapped reads are 17.7X depth. That means (I’ll spare you the math here) that 28,757,062 bases out of the approximately 3,300,000,000 in build 37 were mapped. Or 0.8% identity (99.2% different) to the human genome. Clearly there is a communication issue here- and I am giving them the benefit of the doubt and stating for the record that I may be ignorant.

C. The authors state that the mitochondrial DNA sequence (allele frequency) is conveniently similar to a haplotype group found on the west coast of South America. They provide no evidence to support this.

OK, so let’s circle up here. The authors use the points above to show that the organism is human. ALL of the analysis they do in the paper is predicated on this argument.

The sequencing they do in the main article is the kind of sequencing one would do for a human clinical sample. It is not the kind of sequencing you would do if you didn’t know what you were sequencing. The two are significantly different. Why did Gary sequence the organism’s sample as if it were a human clinical (research) sample? Easy- because that is the kind of sequencing he knows how to do (way way way easier to do) and it fit the vast array of informatics tools and databases available to him. If the only tool you have is a hammer, then every problem looks like a….

His first mistake was using a human reference genome as an assembly scaffold for the raw sequence reads. This needed to be done reference-free, de novo. Steve, use these words when you talk about this: “reference-free genome assembly”. This is a very challenging mathematical task and would require very special compute infrastructure to do. It can be done, but not by a guy like Nolan. Or Butte for that matter. Bustamante, maybe.

Next, they should’ve used multiple different sequencing platforms with different approaches… paired-end (they used), mate-pair, combinatorial, single-molecule sequencing. These methods in conjunction could help generate a complete de novo assembly. From that assembly one could really determine whether or not Atacama is human.

There are a host of fundamental questions that need to be answered from a correct starting point:

• How many chromosomes does Ata have?

• What are the sizes of the chromosomes?

• Is Ata diploid? Triploid? Tetraploid? Haploid? Polyploid? Mixed ploidy? …interesting and important note to this point, Nolan threw out ~360,000 variants that mapped as triploid. The human is diploid. Typically, in a normal human genome assembly we would expect <200 triploid calls. That’s right 360,000 triploid alleles in Ata vs 200 triploid alleles in a normal person. He just threw all that to the side. Frankly, I am surprised he reported it at all… …although I guess he was guarding against the intrepid investigator who would re-analyze the sequence and find this glaring issue.

• Do the reads assemble into genes? Is there a detectable intron-exon structure? Are there trinucleotide codons?

• Do the genes have sequence homology to highly conserved terrestrial genes?

I could go on forever. These guys failed as scientists and the academic community loves them for it. The whole point of the scientific method is to avoid bias and let objective reality show us the truth.