Author Responses to Referees 2

Referee #1 (Remarks to the Author):

Comments re: The sasquatch genome project

1. Firstly - I reviewed a previous version of this ms and am pleased to see that the authors have wholly restructured their ms (it is now substantially improved, with many tangents and irrelevant areas of discussion removed or substantially modified). The previous version was rather opaque - was it dealing with sasquatch or not? I am pleased to see that this new version is bolder and more direct. However, does it present satisfactory data, and include satisfactory analysis of that data?

Given that much of the analysis and discussion concerns genetics, and given that I'm not a geneticist, my general feeling is that any decision about the fate of this ms must fall 'into the lap' of geneticist reviewers. Based on more general issues discussed here, I conclude that the work currently suffers from some areas of obfuscation that prevent me from understanding exactly what it is the authors are proposing. I therefore cannot recommend publication at the moment: if relevant geneticists argue for rejection of the ms, I would concur with this opinion. I admit to being greatly intrigued by the data being presented, however, and (as per last time) wonder if there is a 'Nature-worthy signal' in here somewhere.

The discussion early on of "high quality complete genomes" and so on sounds impressive, and I hoped to be impressed by the data collected and analysis of it. However, right out of the gate (1st line of abstract) we are told that the tissue samples concerned "were obtained... from...

Sasquatch". This is problematic - the authors should state that the tissue samples were hypothesised to be from sasquatch, since the tissue samples were not collected directly from animals themselves: the project is testing the hypothesis that the samples come from an unidentified hominid.

Authors' Response:

Though a number of the samples were taken immediately after eyewitness sightings of Sasquatch, we did change the verbiage to correct this, Line 36, "hypothesized to be" from Sasquatch.

2. The comment, also appearing early on in the ms (p. 4), that the DNA evidence reveals "human DNA interspersed with sequence [data] that is novel and distantly related to primates" raised my eyebrows - what do the authors mean by this? That sasquatch, if imagined to be verified as real, is a hybrid of human and non-primate ancestry? Non-primate? I was interested in seeing this qualified: does it stem from a misuse of the term 'primate'? (I think we all agree that humans are primates).

Authors' Response:

Line 59: Corrected to read: human DNA interspersed with sequence that is novel but primate in origin.

3. It is further stated on p. 27 that phylogenetic results "indicate distant relationships with multiple primate sequences". Again - what exactly do the authors mean here? This is never

explained and I regard it as a major problem - the universally favoured hypothesis (as made clear elsewhere in the present ms) is that sasquatch (if real) is a primate, and specifically a hominoid, and presumably a hominid, a hominine and perhaps a hominin, so the text creates the impression that a non-primate ancestry is being preferred. Or, again, is this because 'primate' is being used to mean 'non-human primate'?

Authors' Response:

We omitted the word distantly in Line 584. The novel primate lineage comes off in the non-human range even though there are human sequences interspersed in the genome. So, we were referring to non-human primate.

4. I was hoping that this would be clarified by examination of the phylogenetic/gene trees. While I'm pleased that several phylogenetic/gene trees have been included, they desperately need redesigning, since the text on the branches is so small that it can't be read without exceptional magnification. As such, the trees are essentially useless.

Authors' Response:

We have attempted to enlarge them but in PDF, they can be enlarged in Reader. We are still trying to make them easier to read and currently have them at a publisher attempting to enlarge them without changing the actual data and distance tree.

4. I appreciate that the authors have apparently tried to include as much data as possible in this

ms. I would say that the photographic evidence (Fig 4A and 4B) should not have been included, since those images are poor in quality, do not appear compelling, and weaken the credibility of the ms. Scientific names are written incorrectly on p. 23.

Authors' Response:

We removed the still of 4A and relabeled 4B as Figure 4. Supplementary video 1 shows the same Sasquatch in greater detail and in high definition and therefore supports the use of Figure 4, however if requested, we will remove 4 also. Since this individual is the donor of Sample 37, we felt that tying her to a photo and video was important. We will remove Figure 4 if the editorial board or the reviewers believe that it should be removed after this explanation. Lines 584 through 586 were corrected as far as the scientific names.

Referee #2 (Remarks to the Author):

Short list of some scientific problems:

1. Repeated and abusive use of vague terms such as "multiple", "numerous", "many" or strange scientific terms such as "anomalous", "strinking", "unexplained" or "unexpected".

Authors' Response:

The words “multiple”, “numerous” and “many” were either removed or greatly reduced in their usage. "Anomalous", "striking", "unexplained" and "unexpected" were completely removed from the manuscript.

2. There is a complete lack of statistical evidence or referenced scientific work for the hair analysis; claiming "they were strikingly distinct from human hairs" without any reference or evidence beyond this bold, authoritative statement is not acceptable.

Authors’ Response:

References 15-19 refer to the hair analysis. Figure 5 illustrates the differences between Sasquatch hair and human hair. Materials and Methods also discuss the methods utilized for hair analysis. Statistics are not utilized in forensic hair analysis.

3. The vagueness of the extraction procedures is also surprising (how the contamination was minimized or DNA recovery maximized?): "The DNA was extracted in a clean room using forensic science procedures that minimized contaminant DNA in the samples while maximizing DNA recovery."

Authors’ Response:

The Material and Methods were supplemental due to the length of the paper and it was stated in the manuscript. The extraction techniques were discussed at length in the Supplemental

Materials and Methods.

4. Some paragraphs are confusing and again, the use of terms such as "clearly" or "as expected" is strange. "Control DNA was obtained from the majority of the submitters and was profiled using Promega PowerPlex16 20. As expected, all submitters yielded complete profiles. In contrast, when the hominin samples were tested using Promega PowerPlex16, partial profiles were evident in almost all cases. These data clearly showed that the samples were not contaminated by their submitters." (how?)

Authors' Response:

The words clearly and as expected were removed. Lines 265 and 266 were edited "These data showed that the samples were not contaminated by their submitters since all of the submitters' profiles excluded as being a contributor to the hominin profiles/partial profiles." to show how the samples were not contaminated by the submitters.

5. The authoritative and unsupported statements follow along the paper: "In the experience of this laboratory, hair samples constitute a very reliable source of DNA."

Authors' Response:

We had previously had the following in the paper but when we revised it, we discarded it due to the increased length of the manuscript, however, we have reentered it into the manuscript, Line 280-283: "Almost all large animal breed registries utilize plucked hair samples as their primary source of DNA for their parentage testing programs. One horse registry alone processes over

100,000 hair samples per year. These hair samples are archived and are viable for many years after they are submitted. Therefore the hominin samples were well within our expectations as far as nuclear DNA yield and quality".

6. Some methodologies are totally unspecific and uninformative: "After extraction, yield gels with 3 L of the extracted DNA were utilized to determine if there was DNA present and whether it was degraded".

Authors' Response:

The Material and Methods discuss this and it is visualized in Figure 7. Yield gels are a well established method for visualizing not only the quantity of DNA in an extraction but also the quality of the DNA since DNA will "smear" on a gel instead of giving a distinct band if the DNA is degraded. Mild degradation is depicted in Figure 13 with the human control that had been purposely degraded prior to extraction.

7. More vagueness (we don't know how many samples are all): "All of the screened samples revealed 100% human cytochrome b and hypervariable region 1 sequences".

Authors' Response:

Line 317 changed to "All 110 screened samples".

8. More methodological problems; in degraded DNA, artifactual amplifications are expected and potentially undescribed alleles have to be sequenced to see what's in the PCR product.

"PowerPlex16 amplification of the hominin samples yielded only partial profiles with off-ladder alleles while amplification of DNA." ents : "Once it was established that the Sasquatch nuclear DNA did not conform to human DNA.."

Authors' Response:

The DNA was not degraded as we had a yield gel of the raw DNA showing clear bands (Figure 7) not smears. As some of us are forensic scientists, we are expert in determining DNA quality and mixture (contamination) interpretation and the interpretation of PP16 which we use in court for our DNA profiles. We included references which support our statement as well as the peak heights on electropherograms showed adequate DNA where the amel X dropout should not have occurred (Figure 9). We also sequenced amel X and the results are seen in Table 4. References 43-46 address the X Dropout. We can address this exponentially if necessary.

9. More vagueness: how were the complete mtDNA genomes obtained? Apparently the samples were sent to a company, I guess they amplified it in overlapping fragments? of which length? Or were they captured with some enrichment method and posteriorly sequenced with next generation technologies? One cannot say the samples were sent and results sent back and that's it

Authors' Response:

From Materials and Methods: Aliquots of purified DNA from all of the samples in this study, along with human controls to monitor for possible contamination, were shipped to Family Tree DNA. Proprietary methods were used to amplify the mitochondrial DNA genome. The DNA was amplified using 48 sets of human specific primer pairs that overlapped. Extra primers were

developed and utilized in case of failure due to mutation. The amplicons were sequenced on an Applied Biosystems® 3130*xl* Genetic Analyzer.

10. Instead of generating meaningless mtDNA phylogenetic trees, the authors should check the well known mtDNA phylogeny (for instance, at PhyloTree) and tell us exactly the haplogroup, and/or subhaplogroup and haplotypes of each sample. It will be self evident anyway that these mtDNA genomes fell within the modern human variation, no need for any tree.

Authors' Response:

We were asked for the mtDNA phylogenetic trees in the first submission so we furnished them. They were actually important since the same mtDNA sequences were found in the next generation whole genome sequencing as in the original mtDNA sequencing. We furnished trees for both the original individual mtDNA sequencing as well as the sequence pulled from the whole genomes. We furnished a table with haplogroups for the various samples (Table2).

11. Unsupported claims (by the way, allelic dropout is a well known phenomena in degraded samples and still the most plausible explanation): "It is noteworthy that AmelX allele dropout occurs in significant numbers of the unknown samples yet seldom occurs in normal human testing."

Authors' Response:

We have previously addressed that the samples were not degraded. We also furnished references for the X dropout (References 43-46) and the peak height (RFU) of the Y peak precludes

degraded or insufficient DNA as a cause of X dropout (Figure 9). The samples were tested twice with different methods (PP16 and amelogenin only) and the dropout repeated. Amel X was also sequenced and it failed on most of the samples though the human controls provided normal sequence. Three of the authors have seen literally thousands of PP16 profiles from both paternity and forensic applications. Only one author has actually seen one case of X dropout and it was because of mutation, not degradation or low yield DNA. From the paper: The genotyping of the amelogenin locus produced the most consistent results across the samples tested. The DNA samples yielded four types of results: XX, XY, Y and null. The dropout of the X amplicon was the most significant of the findings observed with the STR genotype analysis of Amelogenin. (Figure 9, Supplementary Data 3) This dropout was reproduced in several individual samples and was repeatable both in the multiplex of PowerPlex 16 and the analysis of the STR locus, so it is unlikely to be an experimental artifact due to low quantity or degraded DNA (Table 3). The repeatability and number of samples exhibiting the X dropout is inconsistent with what would be expected with normal human allele dropout[43-46]. It is noteworthy that $Amel^X$ allele dropout occurs in significant numbers of the unknown samples yet seldom occurs in normal human testing.

12. More vagueness (how long? how do you assess pristiness?): "DNA samples that yielded long and pristine sequences".

Authors' Response:

The quality of the DNA sequences can be assessed by viewing the electropherograms. These are available if needed. Also the Q30 score for the whole genomes denoted the extremely high quality of the sequences, Lines 544-558.

13. More methodological problems (the observed -again, vague- results are likely due to the amplification of environmental and degraded DNA yielded unspecific PCR products):"The resulting sequences ranged from totally non homologous matches, not found in Genbankafter multiple BLASTs (including dissimilar sequence BLASTs) to novel SNPs and even failure to sequence"

Authors' Response:

The quality of the DNA was addressed using Figures 7 and 13 and the Q30 scores. The novel sequences and SNPs repeated in the genomes although there was no failure in the whole genomes. Failures in the early testing can be attributed to primer design failing to amplify novel sequence.

14. More methodological problems: the MC1R sequences (which primers?, which length?), come apparently from PCR products that were not cloned (this is an standard procedure while working on ancient DNA samples to ascertain the heterogeneities present in the PCR product and also the original extract). Thus the C to T change observed in two samples could be the result of cytosine deamination, a well known phenomenon in degraded DNA. In any case, what can be deduced from this section (and also from the mtDNA and the MYH16 section) is that this samples are modern human DNA samples.

Authors' Response:

We have previously discounted degradation of the DNA using visualization by yield gel and Q30 scores on the next generation sequencing. Also, this is fresh, contemporary DNA, not ancient DNA. Sample 26 was worked up thoroughly including histopathology that showed fresh tissue with no degradation or bacterial contamination. Since a whole genome was sequenced with fresh tissue and still provided the same results as were obtained in the early testing, the only conclusion is that the genome is novel and high quality. It is no different than any other genome sequenced today with fresh tissue or blood. We have added the primers as Supplemental Data 9.

15. Meaningless experiments and sections (obviously the low performance in the SNP chip is due to low quality DNA): "In an effort to mimic severely degraded DNA that could explain the strikingly low SNP matches obtained, one of the human controls submitted along with the unknown hominin samples comprised non sterile blood that was purposely maintained at room temperature in a moist environment for 4 days in an effort to maximize degradation of the sample. Upon visual inspection, hemolysis of the sample had occurred and bacterial contamination, which often correlates with DNA degradation, was seen. An acrylamide gel was loaded with the degraded human sample to assess the degradation and was visualized with ethidium bromide. Smearing was observed".

Authors' Response:

If the DNA is degraded, it will smear on a yield gel. Yield gels have been used for years in forensics and standard DNA testing to assess DNA quality from fresh specimens. We have

figures of two yield gels showing DNA that is not degraded and two samples that are showing some degradation, including the human control, Figures 7 and 13.

16. More meaningless sections (artifactual or missing bands are common in ancient DNA, this has been known for about three decades now): "Some of the samples appeared to produce normal amplicons that resulted in bands consistent with the human controls. Other samples displayed clear bands that appeared to be of different sizes than those expected of normal human amplicons. Yet others had multiple bands. Still other samples failed to amplify at all."

Authors' Response:

This is fresh DNA and fresh dried DNA and the degradation issue has already been addressed. The same samples would sequence long pristine (via electropherogram) sequences up to 900 bases at other loci consistent with human as well as novel sequences hundreds of bases long.

17. More vagueness (which ones?): "and forensic techniques to ensure that there was no human contamination"

Authors' Response:

See Materials and methods: "Since the presence of normal human DNA contamination of submitted samples was a primary concern throughout this study, all samples were thoroughly cleaned in a manner consistent with forensic testing procedures. In order to further rule out contamination from human personnel and lab workers, samples from submitters and scientists

working with the samples were collected for comparison with the results obtained in the various DNA tests.

Hair samples were then sorted into two groups for extraction at DNA Diagnostics. DNA from those samples containing 5-50 or more single hair roots were selected and the roots clipped into 1.5 mL microcentrifuge tubes. The hair roots were thoroughly cleansed with water and ethanol prior to extraction to remove any extraneous DNA.

Hair roots were placed in microcentrifuge tubes for DNA extraction and ATL buffer (Qiagen) was added. These samples were digested with proteinase K (PK, 20 mg/mL) and dithiothreitol (DTT, 1.0 M) at 56°C overnight, followed by a three-step organic extraction procedure using phenol:chloroform:isoamyl alcohol (25:24:1) with an additional PCI extraction. This process was followed by a butanol wash and buffer exchange/concentration into TE^{-4} buffer (10 mM Tris, 0.1mM EDTA, pH 8.0) using Microcon®-100 ultrafiltration devices (Millipore, Billerica, MA)[92-93].

The remaining unknown hairs with only 1-5 hair roots were sent to the North Louisiana Criminalistics Laboratory (NLCL, Shreveport, Louisiana) for DNA extraction and purification. The roots were cleaned with water prior to digestion. The cleaned roots were digested in ATL buffer (Qiagen), PK, and DTT at 56°C until completely dissolved, which generally was overnight. The DNA in this crude extract was purified using the EZ1® DNA Investigator Kit with cRNA (Qiagen) and eluted into TE^{-4} on a BioRobot EZ1® (Qiagen).

Saliva swabs, blood swabs, and tissue cuttings (10 mg) were placed in microcentrifuge tubes for DNA extraction. The samples were extracted using the above mentioned organic method with the exception that DTT was not used during digestion.

Reference samples, in the form of buccal swabs from submitters who collected the unknown hair and tissue samples, were isolated using 50mM NaOH and heated to 100℃ for 10 minutes followed by the addition of 1M Tris (pH 8.3)[98]. The DNA extracted at DNA Diagnostics was quantified using a Nanodrop™ spectrophotometer (Thermo Scientific, Willimgton, DE). Hair samples sent to the NLCL were quantified by real time PCR using the Applied Biosystems Quantifiler® Human kit on an Applied Biosystems Prism® 7000 Sequence Detection System[99]. Samples that yielded DNA concentrations too low to use in standard testing had their DNA concentration augmented using multiple displacement amplification method per the manufacturer's instructions[86].

DNA was then visualized on a 1% agarose gel to determine DNA quality by loading 3 µL of DNA extraction. Appearance of the bands determined the quality and quantity of the DNA extraction (Figures 7 and 9)".

18. More vagueness: "According to the laboratory, the sequences themselves were of very high quality"

Authors' Response:

We have added information and the summary generated by the HiSeq 2000 Next Generation Sequencer showing that the genomes were of high quality via the Q30 score, Lines 478-492, and removed the statement above since it is no longer necessary due to the generated Q30 scores.

19. More methological problems (environmental, contaminating DNA will produce no matches

at GenBank, this is common in all ancient DNA genomic projects): "Some sequences produced no homology matches when BLAST searched against all primate, human, Neanderthal, Denisova, and other sequences in Genbank."

Authors' Response:

We previously addressed that 1. The DNA is not ancient and 2. The DNA was not degraded or contaminated both by visualization and testing in the case of preliminary findings and by the Q30 scores generated by the HiSeq2000 for the whole genomes.

20. Although they claim to have generated 30x genoms, I understand they have only assembled and analyzed the chromosome 11, they don't explain why.

Authors' Response:

We explained that chromosome 11 is highly conserved in primates. In the paper: "Thus, the selective supercontigs comprised an abundance of neural associated and putative tumor suppressor sequences all of which are highly conserved in primates and humans and clearly establish that the Sasquatch is closely related to humans. The high homology with multiple primate lineages (including but not limited to, chimpanzee, macaques, gibbons and marmosets) and with humans as demonstrated in phylogenetic trees (Figures 19, 20, and 21) indicate that the supercontigs contain highly conserved human and primate gene sequences".

The goal of this manuscript was to prove that there exists an unknown primate living in North America since eyewitness reports consistently describe a creature with the appearance of a

primate. It takes years to analyze a genome so in order to prove that we had a novel primate, it stood to reason that an area of the genome that is highly conserved in primates be the first area of the genome to be investigated. The supercontigs supported the existence of a novel primate using chromosome 11, which was the goal of this manuscript.

21. The obvious explaination for this section is that the sample is a mixture of DNAs or has been contaminated at the Sequencing Service (did they use a tag sequence for this particular project?; was the service sequencing primates?): "the Sasquatch consensus sequence that showed homology to human chromosome 11 reference sequence is distantly related to multiple primate lineages including Homo Sapiens, Pan Troglodytes (Chimpanzee), Macaca Mulatta (Rhesus Monkey), Nomascus Leukogenys (White cheeked Gibbon) and Callithrix Jacchus (Common Marmoset)."

Authors' Response:

The UT core lab only sequences human DNA and this statement has been added on Lines 472-473. We also previously addressed that we have added the Q30 scores for the genomes. These scores absolutely prove that there was not a mixture of species in the whole genomes as explained above. Lines 544-558 from the manuscript: The run summary generated by the HiSeq 2000 next generation sequencer provides scores, Q30, for run quality. Q30 can also be used to determine if there was any contamination (or mixture) found in the samples sequenced. According to Illumina, a pure, single source sample would have an average Q30 score of 85. However, if there was contamination present in the sample sequenced, the divergent sequences would compete against one another causing a contaminated sample to have a Q30 score of 40 to

50. The Q30 scores for the first read for the three genomes sequenced had Q30 scores of 92, 88 and 89 respectively. The second read was slightly lower 88, 84.25 and 83.66, but still in line with the 85 average. The Q30 is the percent of the reads that have the statistical probability greater than 1:1000 of being correctly sequenced. Therefore, not only were the sequences from a single source, but the quality of the sequences were far above the average genome sequenced using the Illumina next generation sequencing platform. The high quality of the genomes can be attributed to the stringent extraction procedures utilized whereby the DNA was repeatedly purified. This ultra-purified DNA also allowed for greater than 30X coverage of the three genomes. The summary of the next generation sequencing generated by the HiSeq 2000 Illumina sequencer is furnished as Supplementary Data 7.

22. Sequencing data should be freely available to the scientific community after publishing (even better, before).

Authors' Response:

We attempted to upload all sequences to GenBank but they refused them (I have the email and it said because we do not have a species name). We asked how to do it and they refused to return our emails or calls). As a result, we attached all sequences as supplemental per Dr. Gee's request.

23. In my view, there conclusions are not supported by the data. What do we have here is likely low quality DNA samples belonging to modern humans that yield some conflicting results in

unspecific genotyping approaches. I suggest also some background contamination in the next generation sequencing from previous primate sequencing projects, again likely influenced by the original low quality of the samples.

Authors' Response:

This is an incorrect assumption as the data was repeatable and of good quality. We have defended our position numerous times (above) and have added new data (Q30 scores) supporting our position that the DNA is not degraded nor is it contaminated.

Referee #3 (Remarks to the Author):

I appreciate the additional work the authors have undertaking trying to meet the concerns raised by me and the other reviewers. However, I am afraid that I do not find the paper worth publishing. The authors still lack explaining in any convincing way how this "new species" carries mtDNA genomes identical to those of modern humans only. I also believe that alternative explanations exist as to why the nuDNA genomes differs from that of contemporary humans e.g. mapping a mixture of animal DNA and human contamination to human reference genomes.

Authors' Response:

We felt that the data should speak for itself so we did not include a theory concerning the lack of novel SNPs in the mtDNA or the presence of the modern human mitochondrial DNA across all

110 samples. We have added a theory to the Conclusions Section in the manuscript, Lines 734-738.

We also previously addressed that we have added the Q30 scores for the genomes. These scores absolutely prove that there was not a mixture of species in the whole genomes as explained above. Lines 544-558 from the manuscript: The run summary generated by the HiSeq 2000 next generation sequencer provides scores, Q30, for run quality. Q30 can also be used to determine if there was any contamination (or mixture) found in the samples sequenced. According to Illumina, a pure, single source sample would have an average Q30 score of 85. However, if there was contamination present in the sample sequenced, the divergent sequences would compete against one another causing a contaminated sample to have a Q30 score of 40 to 50. The Q30 scores for the first read for the three genomes sequenced had Q30 scores of 92, 88 and 89 respectively. The second read was slightly lower 88, 84.25 and 83.66, but still in line with the 85 average. The Q30 is the percent of the reads that have the statistical probability greater than 1:1000 of being correctly sequenced. Therefore, not only were the sequences from a single source, but the quality of the sequences were far above the average genome sequenced using the Illumina next generation sequencing platform. The high quality of the genomes can be attributed to the stringent extraction procedures utilized whereby the DNA was repeatedly purified. This ultra-purified DNA also allowed for greater than 30X coverage of the three genomes. The summary of the next generation sequencing generated by the HiSeq 2000 Illumina sequencer is furnished as Supplementary Data 7.