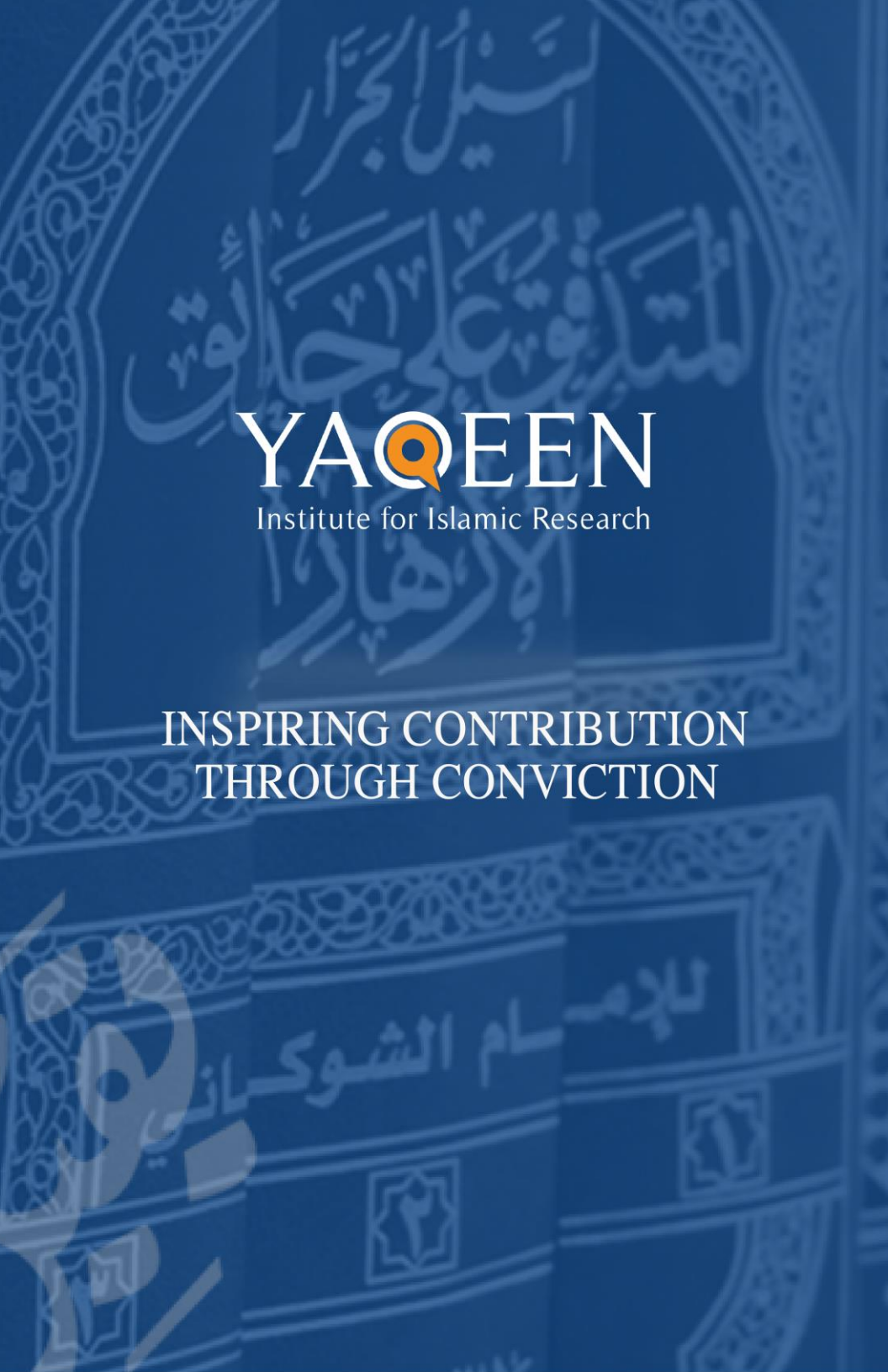




YAQREEN

Institute for Islamic Research

INSPIRING CONTRIBUTION
THROUGH CONVICTION



Can Islam Object to Evolution?

Evaluating Human-Chimp Genetic Similarity

Daniel Haqiqatjou

©2016

Contents

I. Introduction	4
II. “Popular” Science	5
III. Comparative Metrics	6
IV. Remarkable Divergence	8
V. King and Wilson’s 99%	12
VI. Much Ado About 2%	13
VII. Salacious Headlines	14
VIII. Other Studies	15
IX. Draft Sequences	16
X. Selective Comparison	17
XI. The Sequence Alignment Problem	18
XII. “Dark Matter” of the Genome	20
XIII. Conclusion	21

I. Introduction

Evolutionists offer a number of arguments in support of common descent, i.e., the claim that all life descended from a common ancestor. One of the more prominent arguments, an argument that figures heavily in lay conversations and the public consciousness writ large, is as follows: When we look across the animal world and organisms more broadly, we find that there are “striking genetic similarities” between species with otherwise distinct phenotypes (i.e., distinct observable characteristics). For example, despite significant differences in each organism’s external traits, chimpanzee and human genomes are approximately 99% identical¹ (or 98%,² 96%,³ 94%⁴ depending on the research study).

Popular scientists, like Richard Dawkins, often employ claims of human-chimp genetic similarity to further arguments about common descent and to oppose the notion of human exceptionalism and the Christian and Islamic narrative of human origins from a created ancestor, namely Adam.⁵

What is often left unrepresented to the non-specialist public are the details and distinctive nature of this “striking” genetic similarity so often touted by public intellectuals and scientific reporting alike.⁶

Taking a closer look at the scientific literature provides further information that puts these similarity claims into proper context. What is apparent is that the conclusion of 99% similarity is an oversimplification, and the scientific conclusions drawn are much more cautious and, by themselves, much less definitive of common descent than is often assumed in the public discourse.

¹ King, M., and A. Wilson. “Evolution at Two Levels in Humans and Chimpanzees.” *Science* 188.4184 (1975): 107-116.

² Wildman, D. E. “Implications Of Natural Selection In Shaping 99.4% Nonsynonymous DNA Identity Between Humans and Chimpanzees: Enlarging Genus Homo.” *Proceedings of the National Academy of Sciences* 100.12 (2003): 7181-7188.

³ NIH/National Human Genome Research Institute. “Comparing the chimp and human genomes.” N.p., 31 Aug. 2005. Web. 5 Apr. 2014. http://genome.wellcome.ac.uk/doc_WTD020730.html

⁴ Demuth, Jeffery P., et al. “The Evolution of Mammalian Gene Families.” *PLoS ONE* 1.1 (2006): e85.

⁵ “Richard Dawkins– Comparing the Human and Chimpanzee Genomes.” YouTube. 16 Mar. 2012. Web. 5 Apr. 2014. https://www.youtube.com/watch?v=agraHxYui_4.

⁶ Broad Institute of MIT and Harvard. “Comparison of human and chimpanzee genomes reveals striking similarities and differences.” *Broad Institute of MIT and Harvard*. Broad Institute Communications, 31 Aug. 2005. Web. 5 Apr. 2014. <http://www.broadinstitute.org/news/263>.

II. “Popular” Science

When one first hears genetic similarity arguments, it is difficult not to be completely taken in by them. How can anyone argue with 99%? Upon actually delving into the literature, however, one quickly realizes that the issue is not as straightforward as might first appear. For example, Chris Moran, professor of animal genetics at the University of Sydney, remarks:

Depending upon what it is that you are comparing you can say ‘Yes, there’s a very high degree of similarity, for example, between a human and a pig protein coding sequence’, but if you compare rapidly evolving non-coding sequences from a similar location in the genome, you may not be able to recognise any similarity at all. This means that blanket comparisons of all DNA sequences between species are not very meaningful.⁷

Unfortunately, what many fail to understand is that what is found in scientific literature and what is reported to the lay public are sometimes worlds apart, especially when the issue is as ideologically charged as human origins. Complex scientific work gets distilled into soundbites for mass consumption. This is not a problem in itself, but when that filtering process is molded by an ideological narrative such as “cold, hard science vs. irrational Bible thumping,” then that is where simplifications should be reexamined.

With that in mind, what does the scientific literature have to say? What we will find is that comparing two genomes is far from a trivial task. Specifically, a review of the major papers on the topic reveals:

1. All of them *assume* common descent as axiomatic and beyond question. In other words, none of the geneticists researching human-chimp genetic similarity are attempting to prove or provide systematic argumentation for common descent by way of tallying matching nucleotides between two genomes. This is contrary to the popular perception that 99% similarity is an argument, in itself, for common descent.
2. No research study has attempted to compare 100% of the human and chimp genomes in order to determine an overall percent similarity. Each study limits its comparison to subsections of the genome, and, in some studies, including the landmark 1975 paper that first claimed to have discovered 99% similarity, the compared regions constituted less than 2% of the total genome.⁸
3. There is no single agreed upon or widely used metric by which to quantify the similarity of two genomes. In fact, each paper on the topic uses

⁷ ABC Science. “Do pigs share 98 per cent of human genes?” *ABC Science*, 3 May 2010. Web. 5 Apr. 2014. <http://www.abc.net.au/science/articles/2010/05/03/2887206.htm>.

⁸ King, M., and A. Wilson.

a different method and different parameters in selecting and parsing the relevant data.

4. Many of the key assumptions the major chimp-human genome research papers made in determining 99% similarity have since proved to be erroneous.

III. Comparative Metrics

99% of lab mice genes have direct human counterparts,⁹ and 80% of human genes overlap with those of mice.¹⁰ 90% of human-cat genes match,¹¹ and 94% of dog-cat genes match.¹² There is 60% overlap between human and fruit fly genes and 31% overlap between human and yeast genes.¹³

Is 99% human-chimp genome similarity less impressive in light of the fact that domestic cats share 90% of their genes with humans and yeast share over 30% of their genes with us, etc.? What should we make of these various quantitative comparisons? In reality, it is difficult to make sense of these percentages without a uniform metric to reference. Unfortunately, the biological sciences do not provide one.

We must keep in mind that, as of 2014, the gene sequencing that allows for these kinds of comparisons has only been done for a limited number of organisms (cats, dogs, mice, rats, cows, several great apes, fruit flies, yeast, certain bacteria, etc.) and even then, the genomes of very few species have been completely¹⁴ sequenced.¹⁵ For those that have been completely sequenced, only a few have been directly compared with the human genome, such as those of the great apes. So, evolutionary biologists can neither give a robust nor an exact range of similarity, for example, for all mammals, or mammals vs. reptiles vs. fish, or vertebrates vs. invertebrates, or plants vs. animals, etc. This is important because, what if all vertebrates or all mammals fall within an 80%-99% range of genetic similarity

⁹ Gunter, Chris, and Ritu Dhand. "Human Biology by Proxy." *Nature* 420.6915 (2002): 509-509.

¹⁰ Church, Deanna M., et al. "Lineage-Specific Biology Revealed by a Finished Genome Assembly of the Mouse." *PLoS Biology* 7.5 (2009): e1000112.

¹¹ Pontius, J. U., et al. "Initial Sequence and Comparative Analysis of the Cat Genome." *Genome Research* 17.11 (2007): 1675-1689.

¹² National Human Genome Research Institute. "Comparative Genomics Fact Sheet." N.p., 13 Nov. 2011. Web. 3 Apr. 2014. <https://www.genome.gov/11509542>.

¹³ Botstein, D. "Genetics: Yeast as a Model Organism." *Science* 277.5330 (1997): 1259-1260.

¹⁴ "List of sequenced animal genomes." Wikipedia. Wikimedia Foundation, 30 Mar. 2014. Web. 5 Apr. 2014.

http://en.wikipedia.org/wiki/List_of_sequenced_animal_genomes.

¹⁵ "List of sequenced eukaryotic genomes." Wikipedia. Wikimedia Foundation, 4 Feb. 2014. Web. 5 Apr. 2014.

http://en.wikipedia.org/wiki/List_of_sequenced_eukaryotic_genomes.

to each other? If we knew that range, we could make truly comparative statements like, chimp-human genes overlap, say, 50% more than the average degree of overlap between any two other mammalian species. Comparisons like that might at least have a chance at being minimally probative.

The logic here is that we *should* expect a high degree of gene overlap between organisms that are anatomically similar. This is because, in the most basic sense, an organism's phenotype is an expression of its genotype. Therefore, similarities between phenotypes should translate into similarities in genotypes to at least some degree. For example, cats, dogs, chimps, mice, and humans all have similar circulatory systems, gastrointestinal systems, respiratory systems, reproductive systems, immune systems, metabolic systems, and too many other parallels to list. Given this, what percentage of the genotypes should we expect to overlap simply due to all the major phenotypic parallels we observe between two or more organisms? As a rough benchmark, just look at how phenotypically divergent humans and fruit flies are, yet a whopping 60% of our genes overlap!

As a simple analogy, we would not be too incredulous if it were claimed that the technology in an Apple iPhone and a Samsung Galaxy are 99% similar. They are both smartphones of a similar size with similar functionality: making calls, connecting to the internet, supporting applications. There is going to be a high degree of overlap just because these functions require essentially the same hardware: microprocessors, wifi modules, cameras, touchscreens, mics, speakers, etc.

Thus, the claim that the iPhone and the Galaxy are 99% percent alike would not mean much, especially if it turns out that an iPhone and a breadmaker are 60% alike. But if it were claimed that the iPhone and Galaxy are 50% more similar than the average similarity between any two smartphones, then that would imply something significant and unobvious, e.g., either Apple or Samsung is stealing the other's phone design.

In other words, when it comes to human-chimp similarity, is the 99% indicative of something significant about the relation between chimps and humans *or is the 99% simply riding on the particulars of the comparison scheme the researchers chose in determining that figure?* This question is especially crucial given the complex and input-sensitive algorithmic methods used to actually compare two DNA sequences.

Ultimately, genetics and the biological sciences generally do not offer an objective yardstick by which to measure the similarity of two genomes and, in general, there is no straightforward or standard way to give a percent similarity between two multidimensional objects. For example, what is the percent similarity between an apple and an orange? Well, given that there are countless ways to compare the two, a meaningful answer will have to be benchmarked against how similar, on average, we deem other fruits to be to each other, for example.

All in all, the lack of a frame of reference to normalize comparative data renders the 99% similarity factoid essentially meaningless. Multiple renowned research geneticists quoted in *Science*'s, "The Myth of 1%", concur in this seemingly stark assessment:

Researchers are finding that on top of the 1% distinction, chunks of missing DNA, extra genes, altered connections in gene networks, and the very structure of chromosomes confound any quantification of "humanness" versus "chimpanzeness." [...] There isn't one single way to express the genetic distance between two complicated living organisms. [...] Could researchers combine all of what's known and come up with a precise percentage difference between humans and chimpanzees? "I don't think there's any way to calculate a number," says geneticist Svante Pääbo, a Chimp Consortium member based at the Max Planck Institute for Evolutionary Anthropology in Leipzig, Germany. "In the end, it's a *political and social and cultural thing* about how we see our differences."¹⁶

IV. Remarkable Divergence

Chimpanzee and human Y chromosomes are remarkably divergent in structure and gene content.¹⁷

That is the title of a prominent 2010 research paper that adds another dimension to human-chimp genetic comparisons. Hughes, et al., found that the chimpanzee Y-chromosome has only 47% as many protein-coding elements and only two-thirds as many distinct genes as the human Y-chromosome. Also, more than 30% of the chimp Y-chromosome lacks a counterpart on the human Y-chromosome and vice versa (see Figure 1). In one part of the paper, the authors even state: "The difference in MSY gene content in chimpanzee and human is more comparable to the difference in autosomal gene content in chicken and human."

What is truly telling is that Hughes, et al., recreated the DNA comparison of other studies in order to benchmark their alignment techniques.

As expected, we found that the degree of similarity between orthologous chimpanzee and human MSY sequences (98.3% nucleotide identity) differs only modestly from that reported

¹⁶ Cohen, J. "Evolutionary Biology: Relative Differences: The Myth of 1%." *Science* 316.5833 (2007): 1836-1836. Emphasis added.

¹⁷ Hughes, Jennifer F., et al. "Chimpanzee and human Y chromosomes are remarkably divergent in structure and gene content." *Nature* 463.7280 (2010): 536-539.

when comparing the rest of the chimpanzee and human genomes (98.8%).

This means that “remarkable divergence” exists despite the 98% sequence similarity in the Y-chromosome, implying that the rest of the genome may also contain major disparities even though sequence similarity is determined to be 99%, 98%, or 95%.

This is not the first example of divergence that geneticists have uncovered between human and great ape genetic sequences. The field was privy to Y-chromosome misalignments as far back as 1998.¹⁸ Chromosome 4, 9, 12,¹⁹ and, particularly, 21 have also been found to contain “large, non-random regions of difference.”²⁰ Interestingly, these discrepancies are usually investigated and emphasized in research seeking to discover the genetic secret to “humanness,” namely what makes us characteristically human as opposed to mere chimp.

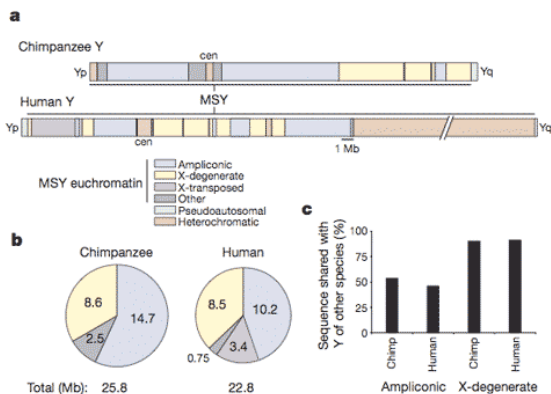


Figure 1 | Comparison of chimpanzee and human Y chromosomes.

a, Schematic representations of chromosomes. cen, centromere; Yp, short arm; Yq, long arm. For both chromosomes, the MSY is indicated. Six sequence classes are shown, four of which are MSY euchromatin. (‘Other’ denotes MSY single-copy sequences that are not X-degenerate or X-transposed.) Chromosomes are drawn to scale, with the exception of the large heterochromatic block on human Yq. **b**, Sizes (in Mb) of four MSY euchromatin sequence classes in chimpanzee and human. **c**, Percentages of ampliconic and X-degenerate sequences present on chimpanzee Y chromosome that are also present on human Y chromosome, and vice versa.

Figure 1

¹⁸ Archidiacono, Nicoletta, et al. “Evolution of chromosome Y in primates.” *Chromosoma* 107.4 (1998): 241-246.

¹⁹ Gibbons, Ann. “Which of Our Genes Make Us Human?” *Science*, 4 Sept. 1998. Web. 6 Apr. 2014. <http://www.sciencemag.org/content/281/5382/1432>.

²⁰ Fujiyama, A., et al. “Construction and Analysis of a Human-Chimpanzee Comparative Clone Map.” *Science* 295.5552 (2002): 131-134.

Discrepancies are also emphasized in phylogenetics, i.e., the genetic analysis used to determine how different organisms are related on the evolutionary tree. For example, in 2007 Ebersberger, et al., claim:

For about 23% of our genome, we share no immediate genetic ancestry with our closest living relative, the chimpanzee. [...] Thus, in two-thirds of the cases a genealogy results in which humans and chimpanzees are not each other's closest genetic relatives. The corresponding genealogies are incongruent with the species tree. In accordance with the experimental evidences, this implies that there is no such thing as a unique evolutionary history of the human genome. Rather, it resembles a patchwork of individual regions following their own genealogy.²¹

One might ask, why have these in depth chromosomal studies, like the one from Hughes, et al., not been conducted for all ape chromosomes? The chromosomes of rodents and fruit flies, for example, are known in great detail, and the reason is those organisms can be experimented on endlessly in medical research. Not so with apes. Ethical standards and animal conservation regulations disallow invasive and terminal experimentation on apes. For this reason, funding for ape chromosome research is relatively sparse because, in the end, there are few opportunities of applications for any findings. Why waste millions of dollars in funding on delving into ape chromosomes when, afterward, one is not allowed, due to ethical standards, to use those findings to further medical science through genetic modification and experimentation on living apes?

In any case, given these known chromosomal and phylogenetic discrepancies across multiple regions of the human-chimp genome map, what are we to make of the 99% human-chimp similarity claim?

The answer lies in the details of the methodologies geneticists use to sequence and align the human and chimpanzee genome. For example, since humans have 46 chromosomes compared to 48 in chimps, is that not a 4.2% difference right off the bat? Obviously, that is deliberately simplistic. But the point is that comparing the human and chimp genomes is not a simple matter of lining the two up and seeing how much they match, though that is precisely the impression a non-specialist may come away with.

²¹ Ebersberger, I., et al. "Mapping Human Genetic Ancestry." *Molecular Biology and Evolution* 24.10 (2007): 2266-2276.

In fact, science and natural history museums with exhibits dedicated to evolution—e.g., the “Explore Evolution”²² project that was featured at numerous natural history museums across the US—often relay²³ this simplistic and ultimately inaccurate notion of genetic similarity to the public by printing a few thousand aligned nucleotides from each genome onto posters side by side, as if to imply that human-chimp genetic overlap is as plain as clear day (see Figures 2, 3, and 4). Just open your eyes and see!

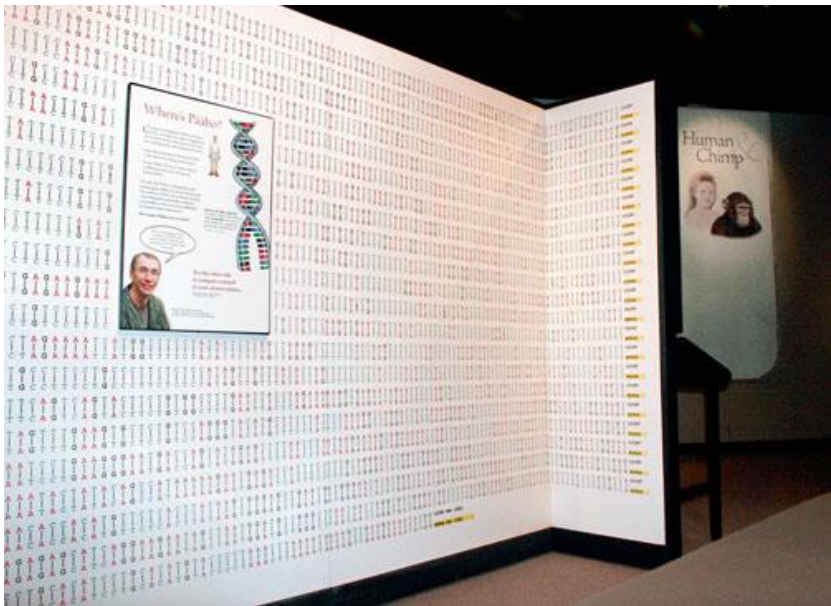


Figure 2

²² Explore Evolution. “Explore Evolution: Project Overview.” University of Nebraska State Museum, n.d. Web. 6 Apr. 2014. <http://explore-evolution.unl.edu/overview.html>.

²³ “Richard Dawkins– Comparing the Human and Chimpanzee Genomes.” YouTube.



Figure 3



Figure 4

V. King and Wilson's 99%

Let's dig into the details of gene sequencing and comparison. The initial research claiming 99% similarity came in 1975 from King and Wilson, who used three biochemical methods to indirectly measure genetic overlap by examining select human and chimp proteins. One important note is that King and Wilson were

not setting out to prove that human and chimp genetics highly overlap. Actually, this was a surprising result for them, and they concluded:

The intriguing result, documented in this article, is that all the biochemical methods agree in showing that the genetic distance between humans and the chimpanzee is probably too small to account for their substantial organismal differences.²⁴

Of course, what was filtered down to the public (and what was interpreted later by many in the scientific community) was that King and Wilson's research provided prime evidence for common descent.²⁵ It is interesting that King and Wilson themselves felt that the discovered genetic similarity belied the vast divergence between the two species, so much so that the role of genetic sequence as the primary determinant of an organism's phenotype was questioned.

VI. Much Ado About 2%

Besides this point, let's also look more closely at King and Wilson's research methods. The first thing to note is that, due to the technological limits of the time, their methods focused on an analysis of human and chimp proteins and not the actual genome. Even then, they only compared a handful of homologous proteins as those are the most readily comparable. Nowhere is it claimed that the selected proteins are representative of the vast variety of proteins in both human and chimp bodies. In fact, King explicitly caveats:

Owing to the limitations of conventional sequencing methods, exactly comparable information is not available for larger proteins. Indeed, the sequence information available for the proteins already mentioned [in this paper] is not yet complete.

Beyond these gaps, what is more significant is that at most, proteins only reflect the *coding* portion of the genome while non-coding areas of the genome are completely missed. Interestingly, 98% of the human genome is non-coding.²⁶

What is the difference between the coding and non-coding regions of DNA? As it is commonly put, DNA carries the genetic instructions used in the development and function of an organism's biology. The mechanics of how these instructions are implemented is quite complex and not fully known, but, to put it simply, the coding portion of DNA encodes the various proteins which serve as

²⁴ King, M., and A. Wilson.

²⁵ Morange, Michel. "The genetic distance between humans and chimpanzees: What did Mary-Claire King and Allan Wilson really say in 1975?" *Journal of Biosciences* 36.1 (2011): 23-26.

²⁶ Meisler, M. H. "Evolutionarily Conserved Noncoding DNA in the Human Genome: How Much and What For?" *Genome Research* 11.10 (2001): 1617-1618.

the fundamental building blocks of bodily function. In humans, less than 2% of all DNA is associated with this coding process.

For decades, biologists have insisted that the non-coding regions of the genome, which constitute over 98% of our DNA, is simply “junk.”²⁷ They reasoned that, since non-coding regions played no discernible part in the formation of proteins, these regions had no biological function. This assumption, of course, has colored all subsequent research on human-chimp genetic overlap.

For King and Wilson’s iconic paper, the fact that their comparison only focused on coding elements of the genome means that the 99% similarity they found is inapplicable to the vast majority—over 98%—of total human-chimp genetic material.

VII. Salacious Headlines

Even if scientific consensus agrees that non-coding regions of the genome play no biological function, it would be a misinterpretation to state that human-chimp DNA is 99% similar based on King and Wilson’s work. As far as King and Wilson are concerned, it would be more accurate to claim, e.g., “Human and chimp DNA is 99% similar... in the 2% of the genome that has been compared.” Of course, a headline along those lines would not attract much attention much less strike anyone as an earth-shattering result.

To make matters worse, the 99% similarity claim of King and Wilson is even less significant once it became apparent²⁸ that “junk” non-coding DNA is not as biologically useless²⁹ as previously assumed. More recently, geneticists are claiming that as much as 80% of non-coding DNA is biomechanically active.³⁰ And, even more strikingly, they are discovering how non-coding DNA plays an essential role in regulating crucial genetic processes. In other words, what was up until as recently as 2010 assumed to be “junk” and was, for the most part, disregarded in comparisons between human-chimp genetics is now understood by biologists to be a critical component of our genotypes. As Ewan Birney of the European Bioinformatics Institute tellingly puts it: “What is remarkable is how much of [the genome] is doing at least something. It has

²⁷ Ohno, S. “So much “junk” DNA in our genome.” *Brookhaven Symposia in Biology*. 23:366-70 (1972).

²⁸ Polavarapu, Nalini, et al. “Characterization and potential functional significance of human-chimpanzee large INDEL variation.” *Mobile DNA* 2.1 (2011): 13.

²⁹ Nowacki, M., et al. “A Functional Role for Transposases in a Large Eukaryotic Genome.” *Science* 324.5929 (2009): 935-938.

³⁰ Park, Alice. “Junk DNA.” *Time*. 12 Sept. 2012. Web. 6 Apr. 2014.
<http://healthland.time.com/2012/09/06/junk-dna-not-so-useless-after-all/>.

changed my perception of the genome.”³¹ Go figure! 98% of our genome is “doing at least something” and is not completely inert waste.

Given the very selective and limited human-chimp genome comparisons that have been done by King, Wilson, and others, it is no surprise that the more focused studies that analyze specific chromosomes in detail, such as the Y-chromosome study cited above, find “remarkable divergences.”

VIII. Other Studies

A review of human-chimp genome comparisons since King and Wilson’s paper shows many of them base their conclusions exclusively on the coding portion of the genome, which only accounts for 2% of the entire genome (e.g., Wildman, et al.,³² Nielsen, et al.³³). The rest of the literature—all of which predates the 2010 research on the importance of non-coding regions—includes both coding and non-coding portions to varying extents (though non-coding regions are generally underemphasized). Nonetheless, these studies limit their comparison to some portion of the total genome, meaning there is no, as it were, end-to-end comparison of the entirety of the human and chimp genomic sequences.

This, of course, is a given for any research that predated the completion of the Human Genome Project (HGP) in 2003 and the chimp draft genome of the 2005 Chimp Consortium.³⁴ Obviously, no one could provide a comprehensive comparison of the entirety of the two genomes prior to them being (nearly) fully sequenced, in 2003 and 2005, respectively. (And even the chimp genome sequence is a draft. More on that later.)

For example, Britten³⁵ in 2002 only compared 846,016 bases out of the total roughly 3.08 billion³⁶ that constitute the human genome, which is just 0.03% of the total. Arnason, et al., six years prior, had only considered 165,000, which is 0.006%.³⁷ Liu, et al., in 2003, compared nearly 5 million, which is 0.17% of the

³¹ Park, Alice.

³² Wildman, D. E.

³³ Nielsen, Rasmus, et al. “A Scan for Positively Selected Genes in the Genomes of Humans and Chimpanzees.” *PLoS Biology* 3.6 (2005): e170.

³⁴ The Chimpanzee Sequencing and Analysis Consortium. “Initial Sequence of the Chimpanzee Genome and Comparison with the Human Genome.” *Nature* 437.7055 (2005): 69-87.

³⁵ Britten, R. J. “Divergence between samples of chimpanzee and human DNA sequences is 5%, counting indels.” *Proceedings of the National Academy of Sciences* 99.21 (2002): 13633-13635.

³⁶ International Human Genome Sequencing Consortium. “Finishing The Euchromatic Sequence of the Human Genome.” *Nature* 431.7011 (2004): 931-945.

³⁷ Arnason, Ulfur, et al. “A complete mitochondrial DNA molecule of the white-handed gibbon, *Hylobates lar*, and comparison among individual mitochondrial genes of all hominoid genera.” *Hereditas* 124.2 (1996): 185-189.

total.³⁸ Ebersberger, et al., in 2002, compared about 3 million, which is 0.1%.³⁹ Anzai, et al., specifically looked at the MHC multigene region of the genome, which is associated with the immune response of vertebrates; in total, it constitutes 0.06% of the genome.⁴⁰ Thomas, et al.,⁴¹ considered 0.06% in 2003 and Nielsen, et al.,⁴² considered 0.6% in 2005.

The only study to take into account a sizable majority of the human and chimp genomes was the 2005 Chimpanzee Sequencing and Analysis Consortium, which compared 2.3 billion nucleotides, i.e., approximately 76.7% of the total.⁴³

In truth, none of these studies unqualifiedly claim 99% similarity between human and chimp genomes. Rather, the caveat is always there (sometimes more explicitly, sometimes less) that the 95%, 98%, or 99% similarity discovered is limited to the *partial* segments of the genome aligned.

IX. Draft Sequences

Now let's dig deeper into modern sequencing and genome comparison techniques in order to get more insight into the findings of the 2005 Chimp Consortium, which came closest to comparing the entirety of the human-chimp genetic sequence.

Prior to actually comparing DNA, geneticists have to first sequence the genomes in question, which is in itself a monumental task. As noted above, only a handful of species' genomes has been completely sequenced. This is because sequencing projects can be expensive. The International Human Genome Project (HGP), for example, required \$3 billion in funding and took approximately 13 years to complete. The 2005 Chimpanzee Sequencing and Analysis Consortium, in contrast, did not attempt to sequence the chimp genome to the same level of rigor as the HGP and only ended up covering 94% of the entirety of the genome.⁴⁴ Rather than sequence the chimp genome all the way to completion, researchers used the human genome as a "blueprint" to assemble isolated fragments of sequenced chimp DNA. This was done under the assumption that humans and chimps are

³⁸ Liu, G., et al. "Analysis of Primate Genomic Variation Reveals a Repeat-Driven Expansion of the Human Genome." *Genome Research* 13.3 (2003): 358-368.

³⁹ Ebersberger, I., et al. "Mapping Human Genetic Ancestry." *Molecular Biology and Evolution* 24.10 (2007): 2266-2276.

⁴⁰ Anzai, T., et al. "Comparative sequencing of human and chimpanzee MHC class I regions unveils insertions/deletions as the major path to genomic divergence." *Proceedings of the National Academy of Sciences* 100.13 (2003): 7708-7713.

⁴¹ Thomas, J. W., et al. "Comparative analyses of multi-species sequences from targeted genomic regions." *Nature* 424.6950 (2003): 788-793.

⁴² Nielsen, Rasmus, et al.

⁴³ The Chimpanzee Sequencing and Analysis Consortium.

⁴⁴ Ibid.

closely related, such that the human genome can be used as a reference to map the fragmented chimp DNA. The overly cynical among us might be tempted to think that the fact that the human genome was utilized to sequence the chimp genome would have important implications for later comparisons of the two.

X. Selective Comparison

The impression the lay public might get from unqualified claims of 99% human-chimp similarity is that geneticists lined up the genomes and compared sequences of the billions of nucleotides constituting DNA structure, i.e., A, T, C, G. For example, here are the first 100 bases of chimp mitochondrial DNA:

**GTTTATGTAGCTTACCCCTCAAAGCAA
TACACTGAAAATGTTTCGACGGGTTTAC
ATCACCCCATAAACAACAGGTTTGGT
CCTAGCCTTTCTATTAG**

And the first 100 for human mitochondrial DNA:

**GATCACAGGTCTATCACCTATTAACCA
CTCACGGGAGCTCTCCATGCATTTGGTA
TTTTCGTCTGGGGGGTGTGCACGCGAT
AGCATTGCGAGACGCTG**

Given that the entire human genome is on the order of three billion nucleotides and the chimp genome is roughly 10% larger, any notion of “direct” comparison is far beyond consideration. In fact, geneticists employ the help of statistical mathematicians and computer programmers to produce algorithms and software—e.g., BLAST—capable of finding alignments between massive sequences.⁴⁵

Before employing software like BLAST, however, geneticists first pre-select regions of the genome they want to compare. This pre-selection is necessary because certain regions of the human and chimp genomes are *too divergent* to be effectively compared using local alignment algorithms. Regions that are highly repetitive are also excluded (or “masked”) because BLAST and other programs

⁴⁵ BLAST Web Portal. <http://blast.ncbi.nlm.nih.gov/Blast.cgi>. Web. 6 Apr. 2014.

would return “inaccurate” results were these regions to be included. (More on this in the next section.) Ultimately, the final percentage similarity *does not* encompass the excluded regions. In other words, *genome comparison is a measure of similarity in sequences that are already similar enough to be aligned*.

Of course, this kind of limited analysis makes sense for researchers who compare genetic sequences between species ultimately in order to investigate shared genes in making strides in medical science. However, it is clear that this methodology is fundamentally limited in its ability to assay overall similarity in the entirety of two genomes. After all, regions of divergence beyond an arbitrarily specified limit are excluded out of hand. Outside of such examples, it is not clear what deeper scientific utility genome comparison has other than being an arbitrary, highly artificial matching game for the purpose of reaffirming deeply rooted beliefs about the interrelation of humans and apes.

To better appreciate this seemingly controversial statement, it helps to actually see the alignment methodologies in action and hear opinions from notable geneticists.

XI. The Sequence Alignment Problem

A non-specialist may wonder how exactly two nucleotide sequences like the ones above are compared. The answer is, there is no one way to do this. In fact, sequence alignment is a very active field, as researchers debate which sequence alignment algorithms yield the most “reliable,” “high-quality” results.⁴⁶ As Stanford Professor of Computer Science Serafim Batzoglou remarks:

Recently, the literature on basic methodology and tools development has been growing rather than shrinking, indicating that the alignment problem is still not solved. How can that be, after nearly 40 years of research and literally hundreds of available tools?⁴⁷

In actuality, the Sequence Alignment Problem is more of a mathematical problem than a biological one. Furthermore, it is an open problem as no definitive solution exists.⁴⁸ Nonetheless, the central problem is easily stated: Given two (or

⁴⁶ Polyanovsky, Valery O, et al. “Comparative analysis of the quality of a global algorithm and a local algorithm for alignment of two sequences.” *Algorithms for Molecular Biology* 6.1 (2011): 25.

⁴⁷ Batzoglou, S. “The many faces of sequence alignment.” *Briefings in Bioinformatics* 6.1 (2005): 6-22.

⁴⁸ Carrillo, Humberto, and David Lipman. “The Multiple Sequence Alignment Problem in Biology.” *SIAM Journal on Applied Mathematics* 48.5 (1988): 1073.

more) sequences of letters (e.g., A, C, T, G) of a given length, how can we quantify the “distance” or “similarity” between them? For example, consider the below sequences:

TCCCAGTTATGTCAGGGGACACGAGCATGCAGAGAC

AATTGCCGCCGTCGTTTTCAGCAGTTATGTCAGATC

This is precisely the kind of data analyzed in the relatively new field of bioinformatics. Without applying constraints, there are exponentially many ways to align the two sequences (two possibilities shown below):

```

--T--CC-C-AGT--TATGT-CAGGGGACACG--A-GCATGCAGA-GAC
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
AATTGCCGCC-GTCGT-T-TTCAG----CA-GTTATG--T-CAGAT--C

          tccCAGTTATGTCAGggggacacgagcatgcagagac
          |||||
aattgccgccgctcgTTTTcagCAGTTATGTCAGatc

```

Gaps, represented by dashes, are considered an acceptable method to align the sequences because, in evolutionary terms, the gaps represent insertions or deletions (i.e., “indels”) of nucleotides in the genetic sequence. Technically, single substitutions and inversions are also allowed, which further expands the space of possible alignments.

Now, given the two possibilities above, which alignment is correct? Out of the large number of possible alignments for this relatively short sequence, how can we determine which alignment *correctly represents* the phylogenetic relation between the two species? After all, we do not have access to the hypothesized common ancestor’s DNA to compare, contrast, and grade each possibility. Be that as it may, once we decide which alignment is correct, we can then tabulate the percentage similarity by counting matches.

The significance of all this is that *the overall percentage similarity of the two sequences ultimately depends on the alignment scheme one chooses*. Furthermore, *the lack of a standardized alignment scheme renders comparative studies across different genomes problematic and the results dubious*.

Large scale comparisons between genomes typically prefer “local alignment” as opposed to “global alignment.” In the example above, the top alignment represents a global alignment and the bottom one, local. In comparative genomic studies like that of the 2005 Chimp Consortium, local alignment is preferred under the assumption that, given long stretches of DNA, only some portions are related in a sea of uninteresting nucleotide sequences. For example, in the local alignment above, this is considered 100% similarity. The non-aligned areas are simply disregarded. This is how alignment using BLAST works; the program takes a sample query of a given length and scans the database genome until it returns all possible matches, some of them of greater or lesser similarity due to indels, substitutions, etc. The relative location of the matches within the context of the whole genome is not factored because, again, the assumption is that the matching sequences are surrounded by insignificant regions whose exact order does not matter. As American biochemist Russell Doolittle notes: “The underlying message is that one must be alert to regions of similarity even when they occur embedded in an overall background of dissimilarity.”⁴⁹ The background dissimilarity, of course, is excluded from the overall percentage similarity calculation.

In truth, when it comes to the human genome, even modest studies have to consider many kilobases of sequence data. Geneticists reduce the complexity of the alignment problem by limiting their sequence comparison to areas that are most amenable to alignment in the first place, which, most fortuitously, also just happen to be the areas of most genetic interest (at least prior to discovering the importance of non-coding, high-repetition regions, transposable elements, etc. by 2010). There are separate computer programs—e.g., DUST—that “mask” these unwieldy, “uninteresting” background regions of the genome.⁵⁰ As mentioned above, masked regions are excluded from the overall percentage similarity calculation. But how significant is this exclusion?

XII. “Dark Matter” of the Genome

To understand the scale of “low complexity” repetitive regions, we can begin by quoting, in full, a passage from a 2011 study by Koning, et al.:

Eukaryotic genomes contain millions of copies of transposable elements (TE) and other repetitive sequences. Indeed, approximately half of the sequence content of typical mammalian genomes tends to be annotated as TEs and simple repeats

⁴⁹ Doolittle, Russell F. “Searching through Sequence Databases.” *Molecular evolution: computer analysis of protein and nucleic acid sequences*. San Diego: Academic Press, 1990. 99-110.

⁵⁰ Morgulis, A., et al. “WindowMasker: window-based masker for sequenced genomes.” *Bioinformatics* 22.2 (2006): 134-141.

by conventional annotation methods. By contrast, only about 5–10% of mammalian and vertebrate genome sequences comprise genes and known functional elements. The remaining 40–45% of the genome is essentially of unknown function, and is sometimes referred to as the “dark matter” of the human genome. The origins of this “dark matter” fraction of the genome have presumably been obscured, in part, by extensive rearrangement and sequence divergence over deep evolutionary time. Understanding the content and origins of this huge uncharacterized component of the genome represents an important step towards completely deciphering the organization and function of the human genome sequence.⁵¹

Transposable elements are DNA sequences that can change position in the evolution of the genome. Prior to studies like that of Koning, et al. and Bucher, et al., in 2011 and 2012 respectively, which proved the importance of transposable elements, TEs were seen as “parasites of the host genome” whose only discernible function was to obfuscate the regions of the genome geneticists were most keen to investigate.⁵² As we have seen, these regions were masked in the historical genome alignment studies, but as Koning, et al., propose, these regions constitute upwards of 66% of the genome. Other estimates range from 40%⁵³ to 50%.⁵⁴

What this means is that studies like the 2005 Chimp Consortium that masked repetitive regions and disregarded transposable areas that surround aligned sequences have excluded up to 40% of the entire genome in their analysis. In 2005 and as late as 2010, these exclusions could be justified on the basis that these regions had no functional significance to the organism and to phylogenetic considerations generally. But, as we have seen, recent research within the past six years shows that such assumptions were gravely mistaken.

XIII. Conclusion

Much more can be said about the scientific details of gene sequencing and genomic comparison. The fields of bioinformatics and evolutionary genetics have been in a state of rapid development over the past decade and show no sign of slowing. As popular media report on these developments, it becomes ever more

⁵¹ Koning, A. P. Jason De, et al. “Repetitive Elements May Comprise Over Two-Thirds of the Human Genome.” *PLoS Genetics* 7.12 (2011): e1002384.

⁵² Bucher, Etienne, et al. “Epigenetic control of transposon transcription and mobility in *Arabidopsis*.” *Current Opinion in Plant Biology* 15.5 (2012): 503-510.

⁵³ Varki, Ajit, and David L. Nelson. “Genomic Comparisons of Humans and Chimpanzees.” *Annual Review of Anthropology* 36.1 (2007): 191-209.

⁵⁴ Shapiro, James A., and Richard Von Sternberg. “Why repetitive DNA is essential to genome function.” *Biological Reviews* 80.2 (2005): 227-250.

crucial for commentators to include caveats and context when translating scientific findings to the lay public. This care and due diligence will help ensure that scientific data is not sloppily misappropriated in buttressing ideological conclusions.

Beyond the perils of slipshod reporting, a recurring theme in reviewing the comparative genomics literature is that the science itself is far from conclusive. For example, multiple major assumptions made in 2005 by the Chimp Consortium study, such as the relevance of non-coding, repetitive, and transposable regions, were unceremoniously overturned by 2010. Yet, it was only through such selective evaluation of the genome guided by these erroneous assumptions that the similarity percentage of 95% was obtained.

What does all this mean for common descent? What is apparent to many specialists, as cited above, is that attempting to quantify genome similarity is ultimately a silly, meaningless endeavor. Hopefully, this essay has provided adequate substance to that conclusion. In the end, declaring 99% similarity by itself hardly factors in favor of common descent, other than sheer rhetorical force in swaying the uninitiated. This does not mean that biologists do not have other perceived evidence for common descent (some of which will be discussed and evaluated in future essays *in shā'allāh*). Darwin, of course, believed himself to have discovered numerous evidences of common descent as well, and without the aid of genetic analysis. None of these other evidences, however, have played a bigger part in the public consciousness and the widespread acceptance of Darwinian common descent than the 99% similarity claim. But, as we have seen, it simply does not live up to the hype.