The Shadow Genome Project: progress report 2

John P. Costella
770 5th St NW Apt 1207, Washington, DC 20001-2673, United States

(September 16, 2025)

Abstract

I resume my analysis of the number of bases in the diploid human genome, and refine my previous estimates: Sally's and my genomes are about 82% longer than the reference human genome, at about 11.5 billion bases; *i.e.*, the reference genome is missing about 45% of the data in our genomes, or about 5.2 billion bases.

1. September 10

The small audience at the Zapruder Film Symposium on May 10, 2003, gasped when I said that I wasn't even alive when JFK was assassinated. I had not shared the world-stopping shock of that fateful day. But almost everyone has their own personal ones. My father unexpectedly dropping dead on October 9, 1979, was my first. If my son Matthew had died on April 23, 1998, you probably would not be reading this. And then another global one, on September 11, 2001.

I have related [1] how a confluence of events happened to have me looking at a television to see the second plane hit the Twin Towers. Live. Of course, it is nowhere near as traumatic as the personal ones. But it still seizes your brain, and changes your life.

And so it was, now a proud American, that when I figured out on August 28, 2025, that I would have something to say about the human genome, I set a deadline for myself of two weeks hence, September 11, 2025—24 years after that fateful day, and 24 days after my shock at realizing that there was something wrong with the data, on August 18. After all, 24 is my number. Our number. Corny numerology, but my life has always been numbers.

On the last day before going live, I decided to bring my personal laptop along to work to finish whipping up the first progress report, in spare moments between doing my real work. Just before 1:00 p.m. I realized that the new program that I had written and the graph that I had created that morning used the exact same philosophy as Bear. Coincidentally, Sally had dressed me in one of the two custom "Bear 2023" Ralph Lauren polo tops that she'd ordered as gifts for me when I released Bear on what would have been my Dad's 100th birthday. I highlighted the sentence I had just written into the report about Bear and took a remote-control selfie of the laptop screen with my back (and "Bear 2023") above it, from a side office, and sent it to her.

During my regular lunchtime walk to the White House, I sent both Sally and Greg Burnham screenshots of Google's "doodle" for the day: *DNA* [2]. What the fuck? I quickly confirmed that it wasn't because it was a special anniversary of Crick and Watson, or anything like that; it was just "to kick off the school year." It was the latest of an exponentially increasing number of synchronicities. What the fucking fuck? *Of course*—we joked—if it was caused by a non-human intelligence, then they made a mistake: they were a day early. Half an hour later, Sally texted, "I wonder [if] that doodle will be up tomorrow as well?" "Probably not," I answered. She gave my answer a thumbs-up.

At 3:38 p.m. she texted me a photo montage of her seeing Ted Cruz at the Heritage Foundation, where she'd managed to shake his hand. Two minutes later she explained that they were both wearing cowboy boots. Fun fun. I was back at my desk, and had just opened my personal laptop again. I glanced up at the TV on the wall about 20 feet in front of me, silently playing CNN, which a public policy colleague needed to monitor.

"CONSERVATIVE ACTIVIST CHARLIE KIRK SHOT DURING EVENT IN UTAH"

I grabbed my phone and took a photo of the TV, and sent it to both Sally and Greg with a "WTF." My opinion of CNN as a reliable news source being roughly equivalent to that of something scrawled on a hamburger wrapper lying on the filthy streets of DC, I checked the internet, and sent a screenshot of Wikipedia—*Wikipedia*, of all things; my brain was clearly already misfiring—that seemed to confirm the report.

I tried to continue with the progress report, but my brain was shutting down. I could feel it. I put new comments and summaries in different parts, and tried to get it done as quickly as possible. But at 4:40 p.m., the dreaded text from Sally: "Oh shit, they've announced he died."

Writing these words, on September 14, my brain is again shutting down.

...

But today, this morning, my cylinders have at least started firing again. Knowledgeable people have already been providing comments and questions about the two papers. (Sadly, one demonstrated their brainwashed ignorance by claiming Charlie Kirk to be a white supremacist, and recommended that I remove the dedication, lest I be canceled as one as well. Dumb fuck.) This morning I'd again started being woken with understandings of how to resolve the issue of the exact number of bases in the diploid genome. I decided I need to push out a second progress report describing it, before our vacation less than two weeks from now.

So that's what this is. No humor in this one. Functioning again comes first. Humor we can think about some other month.

2. Corrections

I intend to include this section in every Progress Report (hencefort, "PR"), starting with this one, to list corrections for any errors that I find in the paper [3] or previous PRs—for now, just PR1 [4]. I'll try to credit anyone who points out the error to me, but right now it's just me finding my own mistakes.

2.1. Simple `grep` *proofs*

On pages 40 and 41 of the paper I showed that you can check with nothing more than your own eyes, your computer, and the NIH's published reference genome whether my claims of particular sequences being or not being in the reference genome are true or not. In doing that direct proof I grabbed the original copy of the file that I had downloaded from the NIH, and copied it to my working directory, to do my `grep`ping.

Unfortunately, I happened to grab what the NIH calls the "Submitted GenBank assembly" file, `GCA_000001405.29_GRCh38.p14_genomic.fna`. There's nothing wrong with that file, but it's not actually the one that I had been using for months. I had sourced my data from the "NCBI RefSeq assembly," namely, `GCF_000001405.40_GRCh38.p14_genomic.fna`. I only realized the mistake yesterday, September 15 (yes, it's two days later now), when double-checking something from PR1 against the "alternative" version of the reference genome data that I had created for the other project [5], and realized that the source FASTA files had differences in the first header line. It actually makes no substantive difference—the sequence `GGACTCTTATAGTTACCA` appears on line 23,199,678 of both files, and `AATCCCAGTGGCGTCATACTGCATA` appears in neither. But to avoid confusion, and because the change is so trivial (not quite a typo, but almost), I waived my normal rule of not making any changes to PDFs of published papers except for typos, and changed the filename to the correct one in the paper. So if you find that the PDF of the paper has changed slightly from the version that you downloaded earlier, that will be the reason.

2.2. The number of clean needles

On page 22 of the paper I explain why the number of bases is about 6.67% lower after cleaning, as it just represents trimming the first 10 bases from every raw sequence, almost all of which have 150 bases; hence, almost exactly 1/15 of the data represents split ends that I trimmed off. On the next page I report the number of clean needles for Sally's data and for mine. I then move straight onto the singles peaks, and say that their reductions are each slightly larger than the 6.67% that I expected. *What?* I don't know why I mentally conflated the number of *needles* with the number of *bases*, on this particular occasion, other than maybe because I was still kicking myself at not having computed Figs. 13 and 14 weeks earlier. In any case, it happened so late in the game that I didn't pick up my own lapse in logic. Stupid.

Of course, the number of *needles* will be reduced by a factor of 10/135, or about 7.41%, because we use up the first 15 bases in initializing our needle photocopier. On page 23 I reported that Sally's total number of clean needles is 390,091,247,764, which is indeed about 7.41% lower than the total number of her uncleaned needles, 421,298,960,365, that I originally reported on page 2. Likewise, mine drops from 384,534,620,798 to 356,049,894,000, again about 7.41% lower.

Against this background, the 7.7% drop in the frequency of her singles peak, and 7.8% drop in mine, are both within tolerances, given the approximate way that I computed the position of each peak (fitting the top of each with a parabola).

So ignore those comments about them being bigger than I expected. I expected wrongly.

2.3. Bias curves

Figs. 1 and 2 of PR1 have several mistakes. Given the events of the day, you can see how error-prone my work is when I don't OCD-proofread it to death over an extended period of time.

The most obvious mistake is that the caption of Fig. 1 is completely wrong. Clearly, I'd just copy–pasta-ed the source code for the figure from Fig. 1 of the paper, and hadn't gotten around to changing the caption. It should have said something like, "The bias in the number of needles in the reference genome compared to those in Sally's data."

Second, the vertical axis of each of Figs. 1 and 2 of PR1 is labeled "Relative representation of that needle in the reference genome," implicitly "with respect to Sally's," but it's clearly the other way around, since the low-frequency noise goes *up*, to a factor of 9, in Fig. 1 of PR1, whereas the reference genome actually *suppresses* this noise. I had flipped and flopped about which way around I wanted to compute the ratio, and had intended to graph what I stated (the reference genome compared to Sally's), but somehow ended up with the wrong version. Anyway, shit happens, and it was the least of my concerns on September 10.

The third mistake—or poor choice—is more subtle, and it doesn't actually change what you will see on the graphs (but if you downloaded the code before my fixes, it will change the numbers slightly). I had computed the *mean* frequency of the reference needles for each given frequency in Sally's data, but it makes more sense to compute the *sum* of the frequencies. It really only changes the normalization (and it *does* make a small difference, even though you might think that the mean is just the sum divided by the number of needles, because that denominator changes from frequency to frequency), but it makes more sense to me to do it this way.

3. The size of the diploid genome

If you have read the paper and PR1, you'll know that my estimate for the number of bases in Sally's diploid genome went from 6.33 billion in Sec. 5 of the paper (amazing that we can measure it just from the raw data!) up to 12.66 billion in Sec. 9 (it's doubled!), up slightly to 12.70 billion in Sec. 10 (after cleaning off those first 10 bases of each 150-base Nebula read), but then down to 11.05 billion in PR1. This last is a 13% drop, and demands some examination.

There are two different reasons for this drop. (Nothing in this project is ever straightforward.)

The first—which I mentioned in passing in PR1—is that my Figs. 1 through 12 and Fig. 15 of the paper show *the number of distinct needles* as the vertical axis. I had equivocated between doing it this way (the normal way that I do frequencies-of-frequencies tables in my day job) or instead graphing the *total number of needles*, which of course just means multiplying every frequency-of-frequency by the corresponding frequency. (I need a better term for frequency-of-frequency, so that all this sounds less confusing, but it's probably too late to discover or invent one now.) Mathematically, it's trivial, and I could (and did) do it in the Excel spreadsheet of the program's output if I really wanted to. (I've since added it as an option to a number of the programs, just for convenience.) I knew that the shapes of the graphs were (obviously) slightly different depending on the choice, but the gross features were not. The number-of-distinct-needles choice just happened to be the way that I was leaning in my exploration of the raw data at the time that I realized that there were far too many singles, and so that arbitrary choice effectively got "frozen in" at that time.

In retrospect, I probably would have gone the other way. But 20:20 hindsight is wonderful. So let's take a look at how some aspects of the paper would have looked if I had.

Let's forget about the data that I had before I "cleaned" it all in Sec. 10 of the paper (I have now renamed all data files before that point with the additional suffix "`-unclean`," so that I don't accidentally use any of them instead of their "`-clean`" counterparts), and just pretend that I had had the foresight to clean it all from the beginning. The total-number-of-needles equivalent of Fig. 1 of the paper (actually cleaned there in Fig. 15) is shown in Fig. 1. (You can tell when I'm referring to a figure in *this* paper because of those red hyperlinks.) Nice: I don't even need to rescale the vertical axis, like I did in Fig. 2 of the paper. (Hindsight.) The total number of needles between any two frequencies is now just the area under the curve between those two frequencies. (More niceness.) What I referred to in passing in the paper as the "infrared divergence" (well, what do you expect from a quantum field theorist?) now actually peaks at a frequency of 3, rather than diverging in the low-frequency limit. The singles peak is now at a frequency of 65.5

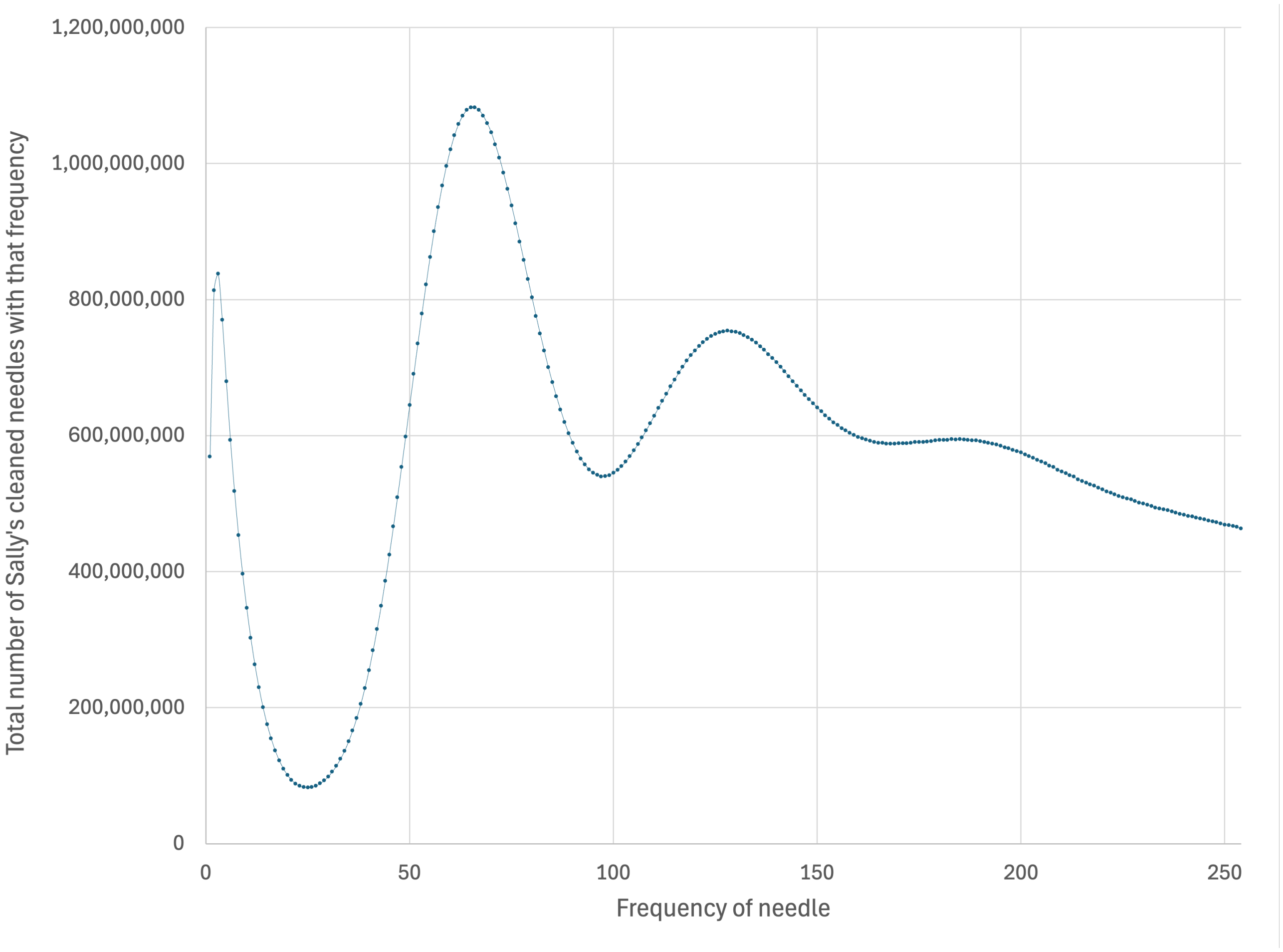

Figure 1: The total number of clean needles for each frequency less than 255 in Sally's data.

(fitting it with a parabola gives me 65.51, but, again, take this precision with a grain of salt, for the moment). The doubles peak is at a frequency of about 128.2.

These numbers sound similar to what I had back at the start of the paper. *But no:* that makes no sense. On page 5 there I had Sally's singles peak at a frequency of 66.5, but that was for the *uncleaned* data. On page 23 it dropped to 61.4 for the cleaned data; this 7.7% drop now (per Sec. 2.3 above) makes sense given the split ends that I trimmed off the raw data. It is this 61.4 that we need to compare with the 65.51 above, for exactly the same data but weighted by the frequency (to turn distinct needles into total needles) before being graphed. In other words, the singles peak has moved *up* by 6.7% (it's purely a numerical coincidence that this is the same percentage that we reduced the raw sequence data by when we cleaned it) by graphing the total number of needles rather than the number of distinct needles. This pushes the length of her haploid genome *down* to 5.95 billion needles, just by itself.

How can graphing the data differently change my estimate so much? Well, as I noted right there on page 23 of the paper, the only way that I could convert the *mode* (peak) of the distribution to a *mean* (needed to estimate the size of the haploid genome) would be if I were able to model the data mathematically. But my attempt to model it with Poisson distributions failed dismally, so I have no precise model. All I have left is my eyeball, and gut feeling, to get a ballpark figure. And that is all that the 6.35 billion or 5.95 billion numbers are—which is why I warned you then (and now) to not put any faith in this extra precision as a *physical* precision (but I *do* retain this degree of precision for *relative* calculations).

So, you ask: am I now saying that my best estimate for Sally's haploid genome is now 5.95 billion bases?

Not yet. I can actually do better.

Let's go back to the paper. I was quite excited to confirm that almost all of the low-frequency needles (then a "divergence," now just a peak) were filtered out if we filtered down Sally's data to only needles seen in the reference genome. That confirmed to me that they almost certainly represent irrelevant noise: sequencing errors or contaminants. Excellent. But *where did I seek to remove that noise?* I didn't. Idiot. So now I will.

Let's start with Fig. 1 itself. I can compute a rough estimate of the number of needles in the low-frequency "noise peak" by just calculating the area under the curve up to, say the first minimum, which occurs at a frequency of 25. The answer is 8,206,282,843 needles, if I include the frequency of 25 itself. That's 2.10% of Sally's 390,091,247,764 clean needles. Not a huge amount, but *they're not Sally's.* Maybe they belong to some bacteria or viruses in her mouth at the time of her cheek swabathon, or maybe they're sequencing errors. It doesn't matter: *they should not be in the denominator.* In other words, I should have considered the total number of *her* needles to be something like 381,884,964,921. That pushes the length of her haploid genome down to 5.83 billion bases.

But I can do even better than this. Let's go back to Fig. 5 of the paper, where I actually filtered out most of the noise from our joint distribution. I could have done the same with each of our *individual* distributions. I didn't. Why not? I don't know. Duh. I got enough from the joint distribution to keep moving forward. But let's do that now. I show the results, for Sally, in Fig. 2. Just as with Fig. 5 of the paper, almost all of the low-frequency noise has been eliminated. The number of needles at the singles peak of 66 (taking the mode of that peak) is 542,157,198, which is very close to half (50.08%) of the corresponding number 1,082,620,440 for Fig. 1, just as we found in the paper when flipping between Figs. 5 and 6. Fitting a parabola to the singles

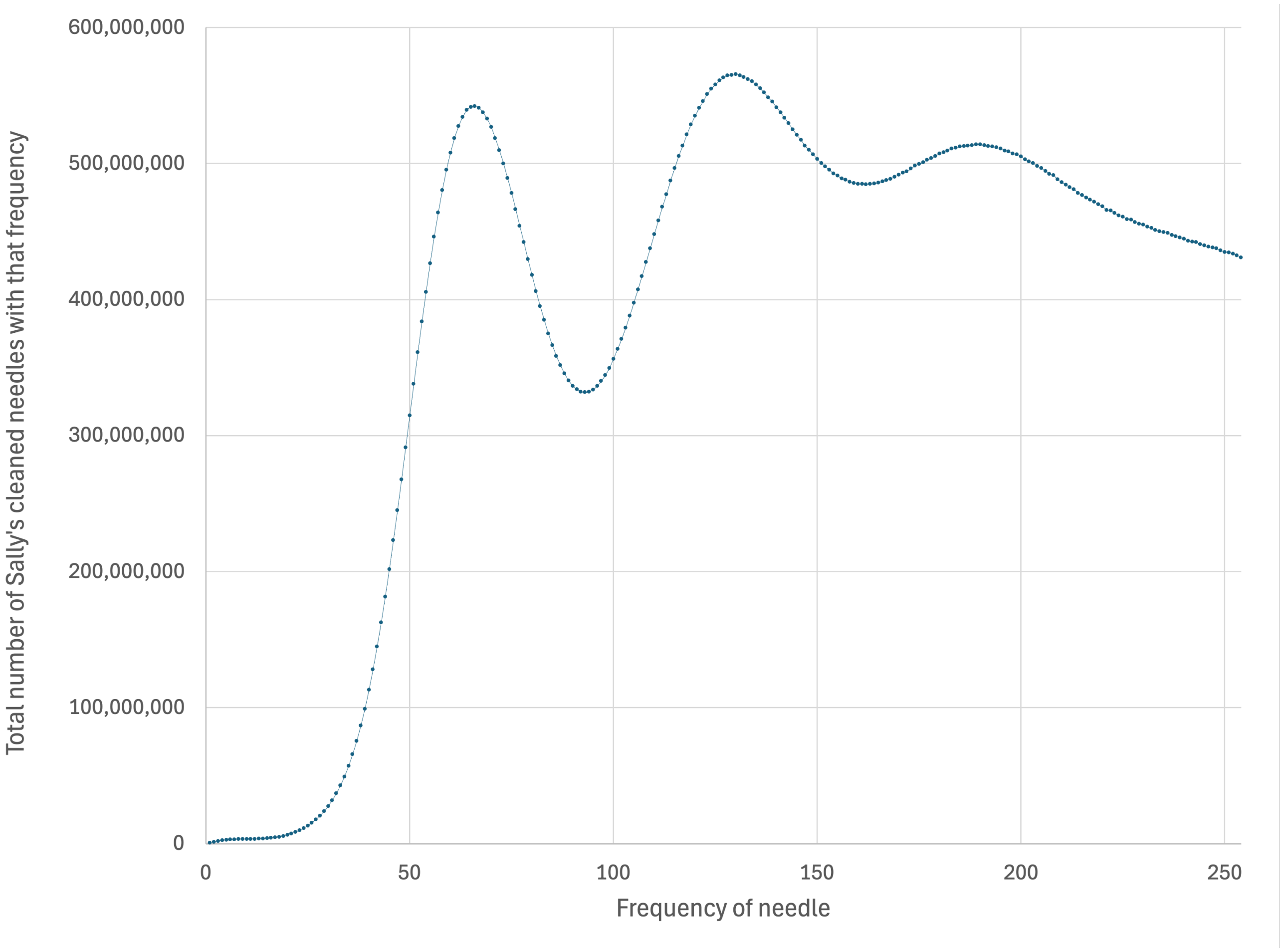

Figure 2: The same as Fig. 1 but only for needles in the reference genome.

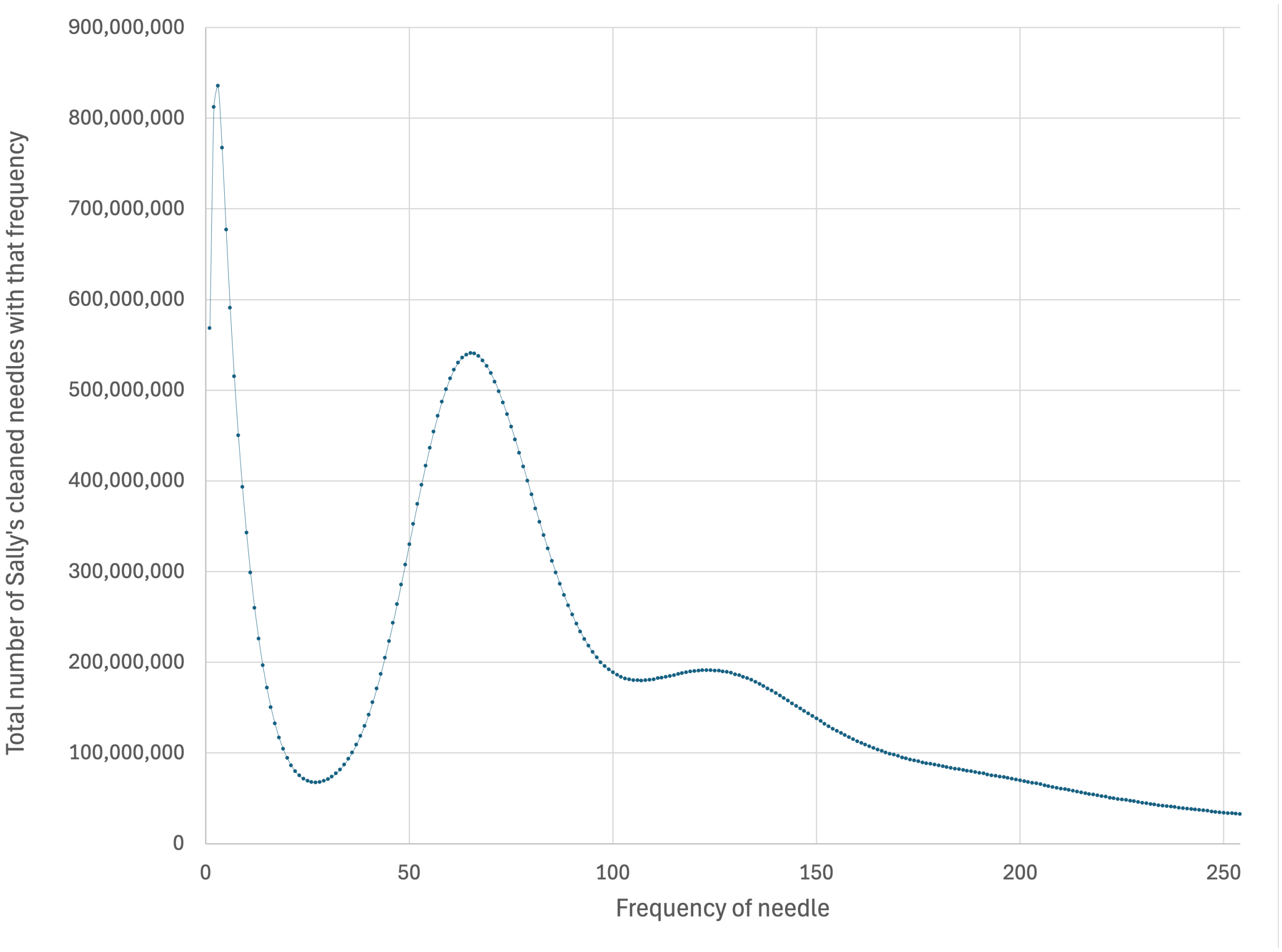

Figure 3: The same as Fig. 2 but now only for needles *not* in the reference genome.

peak, it has now shifted slightly to a frequency of 65.78 (from the 65.51 of Fig. 1), which again emphasizes that the exact position of the peak in any graph is influenced by whatever other "tails" it is added to; here, the low-frequency noise peak has been almost completely eliminated, but the doubles peak has been enhanced (artificially, by the filtering logic, as I described in the paper), and so the singles peak is "sitting on" mainly the tail of the doubles peak, which pulls the maximum up slightly. The equivalent of Fig. 6 of the paper, namely, the equivalent of Fig. 2 for needles *not* in the reference genome, is shown in Fig. 3. Of course, the frequency-wise sum of Figs. 2 and 3 is just Fig. 1. We can see that almost all of the low-frequency noise has been collected in Fig. 3, and, again, the doubles peak in Fig. 3 is artificially suppressed. We expect that these two properties of Fig. 3 will pull the singles peak *lower*, and fitting it with a parabola confirms that it has now been pulled down to 65.28. These variations (up 0.41%, and down 0.35%) are not huge, but they do give us some sort of idea of how uncertain we should consider ourselves to be in locating that singles peak. But it's actually nowhere near as great a difference as moving to the *total* number of needles, which itself is interesting.

The data underlying Figs. 2 and 3 can now give us a rather good estimate of the number of needles in the low-frequency noise peak. The total number of needles (area under the curve) from frequencies of 1 through 66 in Fig. 2 is 10,553,562,025, and for Fig. 3 it is 19,428,026,733.

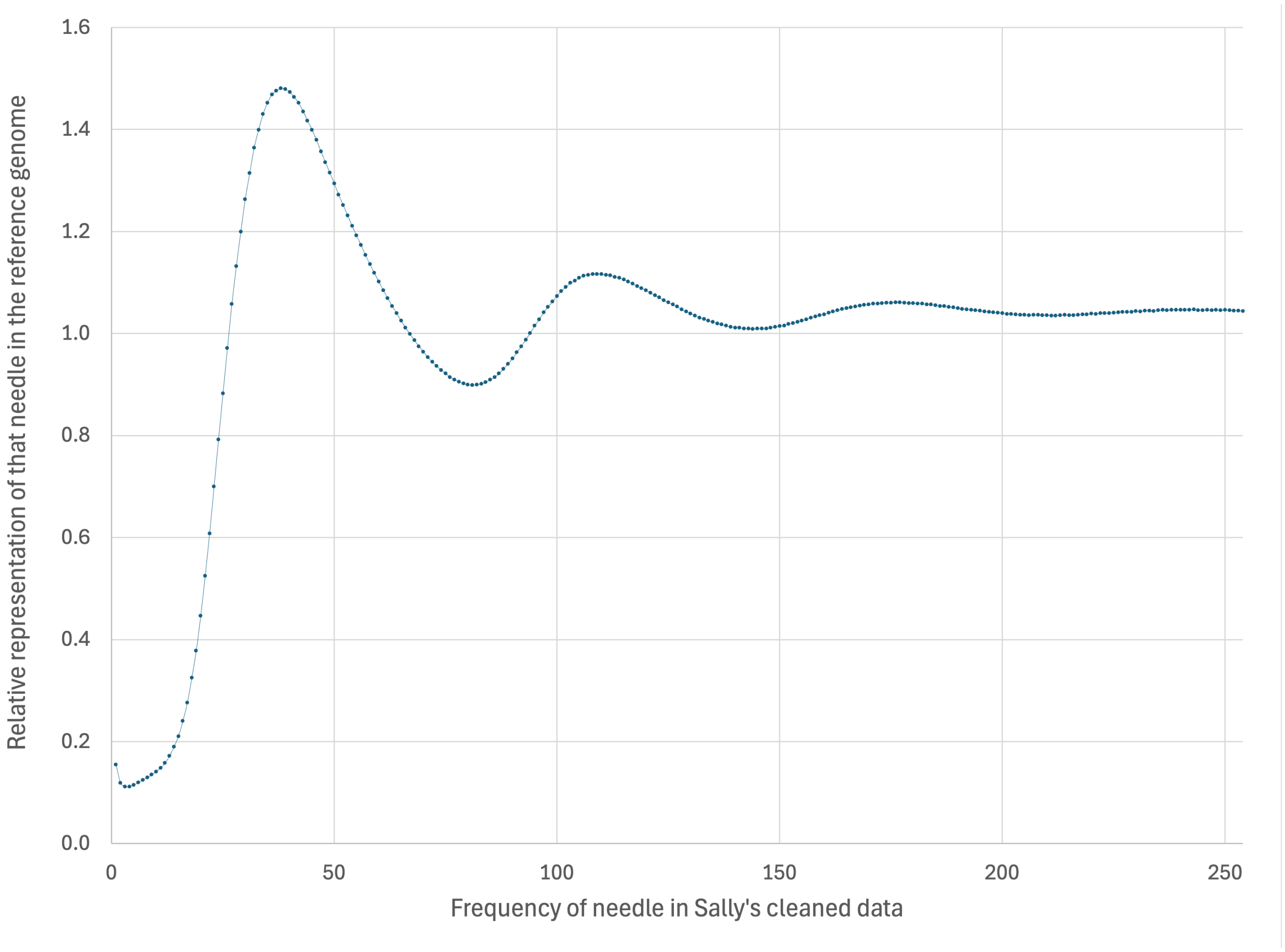

Figure 4: The bias of the reference genome against Sally's clean data. (Fixed from PR1.)

The difference between these, 8,874,464,708, is essentially the noise filtered out by filtering to the reference genome. That's 8.1% larger than the 8,206,282,843 I estimated above, which makes sense because the magic of the reference genome filtering (essentially leveraging the decades of work put into that) selects out even the tail of the noise peak that would otherwise be buried under the singles peak (recall, I only went up to the minimum between them above). These 8,874,464,708 noise needles now represent 2.27% of Sally's total of 390,091,247,764 clean needles. We're left with 381,216,783,056 non-noise needles. Well—maybe not: we can probably clean away the low-frequency part of Fig. 2—say, those with a frequency up to 20—which were so visible in the corner of the origin in the way that I graphed Fig. 5 of the paper. That's another 68,783,555 needles that we can likely consider to also be noise (only 0.018%, to be sure), which leaves us with 381,147,999,501 non-noise needles. If we take the singles peak to be the "middle number" of 65.51 that it is in Fig. 1, then our estimate of the number of bases in Sally's haploid genome is now 5.82 billion, down only slightly from our last estimate above of 5.83 billion.

We're clearly starting to converge in on some sort of an estimate. But we can look at it one more way as well. I noted in Sec. 2.3 above that the bias curves of Figs. 1 ands 2 of PR1 have several errors. I fix them in Fig. 4. You can now see the reference genome filtering out the low-frequency noise, although there is clearly a remainder there. The bias curve then peaks at a frequency (in Sally's cleaned data) of about 38, where the reference genome has almost 50% more frequency than it does, on average, across all needles. What's that about? Well, a frequency of 38 in Sally's data represents *needles that only appear on only one of her two haploid genomes*. Some of the time this will represent what "normal" people have, and the rest of the time it will be a variant. At this point it doesn't matter what the proportions are; what the first peak in Fig. 4 is telling us is that the reference genome has a greater proportion of it than it does for the average needle. This kind of makes sense, because the reference genome is effectively applying whatever choice the compilers of the reference genome have made—as being the non-mutated population choice—to *both* haploid genomes of any particular individual, whereas Sally clearly only has it on either the Edith or the Vic. But if that were *always* the case, then I think the bias would be 2, rather than around 1.5, so I think that sometimes the reference genome has made the other choice. I am still foggy on this, so take that last sentence with a grain of salt.

Then we have a *trough* at a frequency (in Sally's data) of around 81. In PR1 I was thinking of this as being at the singles peak, with the argument being that if the reference genome was biased at 1.5 for needles only on the Edith or the Vic, then it was "taking those needles away" from those appearing on both. But it's now not clear to me how these ratios should be interpreted, and it's even less clear that this trough should appear exactly at the singles peak.

What seems more robust is the fact that the distance from the first peak to the second, in terms of the frequency in Sally's data, is 71; the distance from the second to the third is 67. Likewise, the distance from the first trough to the second is 63, and from the second to the third is 68. Ignoring the fact that the first peak at 38 seems anomalously high—which might be explained by the fact that the data drops to nearly zero on the left but is around unity on the right, again pulling it to the right—it seems that the these values are not far from the estimate of 65.51 of the singles peak above. I'm willing to declare that my understanding of this bias graph is too weak to draw as firm a conclusion from it as I did in PR1, and simply note that it at least seems consistent with the analysis above.

What also struck me about Fig. 4 (or its equivalents in PR1) is that for "large" frequency (here around 250) the oscillations don't asymptote around unity, but rather around a value around 1.04. When compiling PR1 I thought of this as balancing up the "hole" at low frequencies from the reference genome filtering out the noise. At face value this might represent up to 4% noise, whereas I only got a total of about 2.29% above. But if I graph the bias curve out to higher frequencies in Sally's data—say, 1000, as shown in Fig. 5—then I see that this value of about 1.04 is not an asymptote; it decreases towards unity. (It has dropped to just over 1.02 at the right side of this graph.) I think there is some way to compute an asymptotic property to determine the true noise-filtering percentage—probably from cumulative distributions—but I have not yet put my finger on the correct formulation.

Given this lack of clarity, I'm prepared to back away from trying to use the bias curve in the way that I did in PR1—at least for now—and to withdraw the corresponding estimate of 11.1 billion bases for Sally's diploid genome.

That leaves me with twice the 5.82 billion bases that I estimated above for her haploid genome, or about 11.64 billion bases. (Again, I don't claim that this precision is physical; I keep this number of significant figures only for comparison purposes.) This is still significantly below my original estimate of 12.7 billion bases. I will discuss this further in the next section.

Before that, however, let me do the same for my data. The equivalent of Fig. 1 is Fig. 6. In my case the first trough is at a frequency of 21, and the total number of needles with frequencies

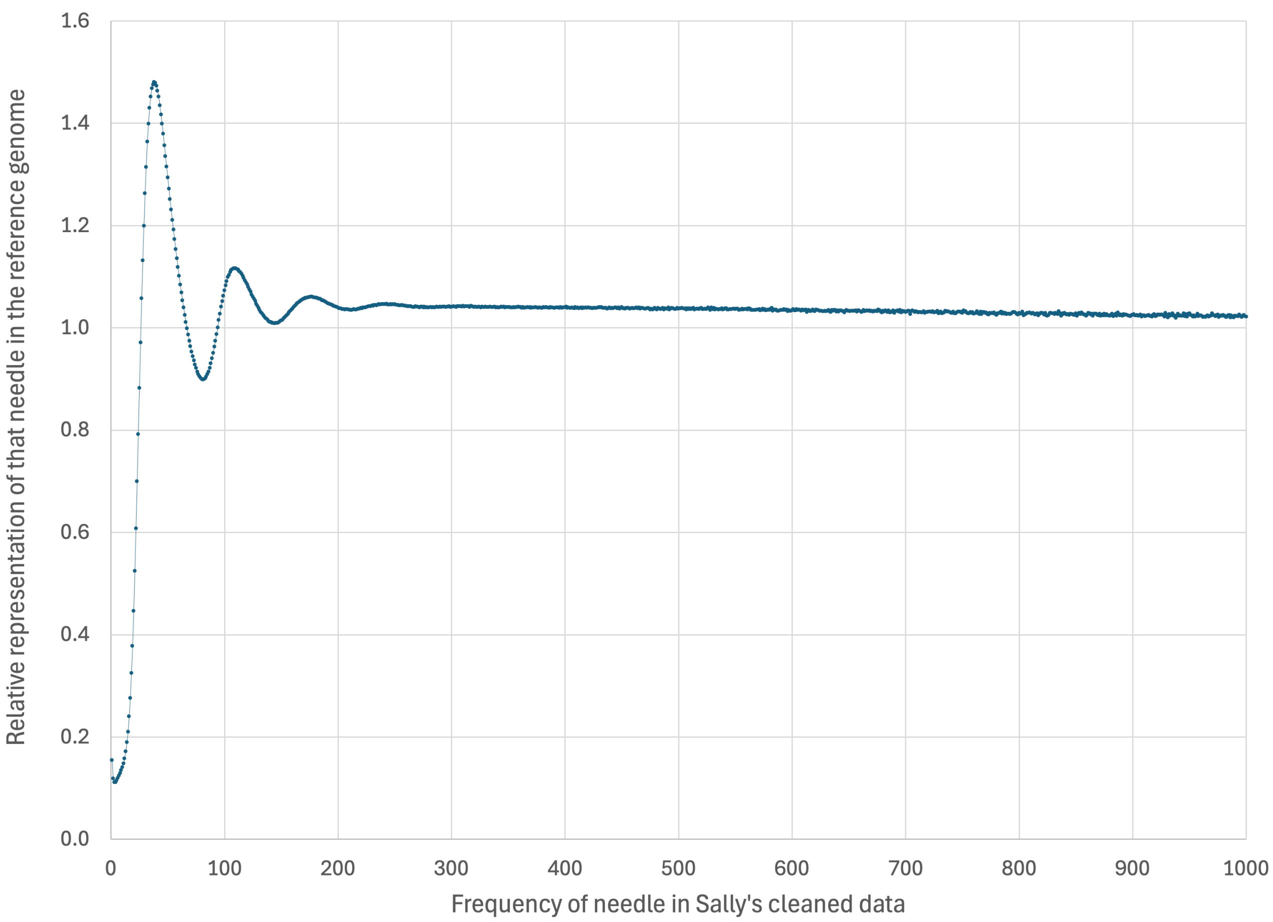

Figure 5: The bias of the reference genome against Sally's data for a greater domain of frequencies.

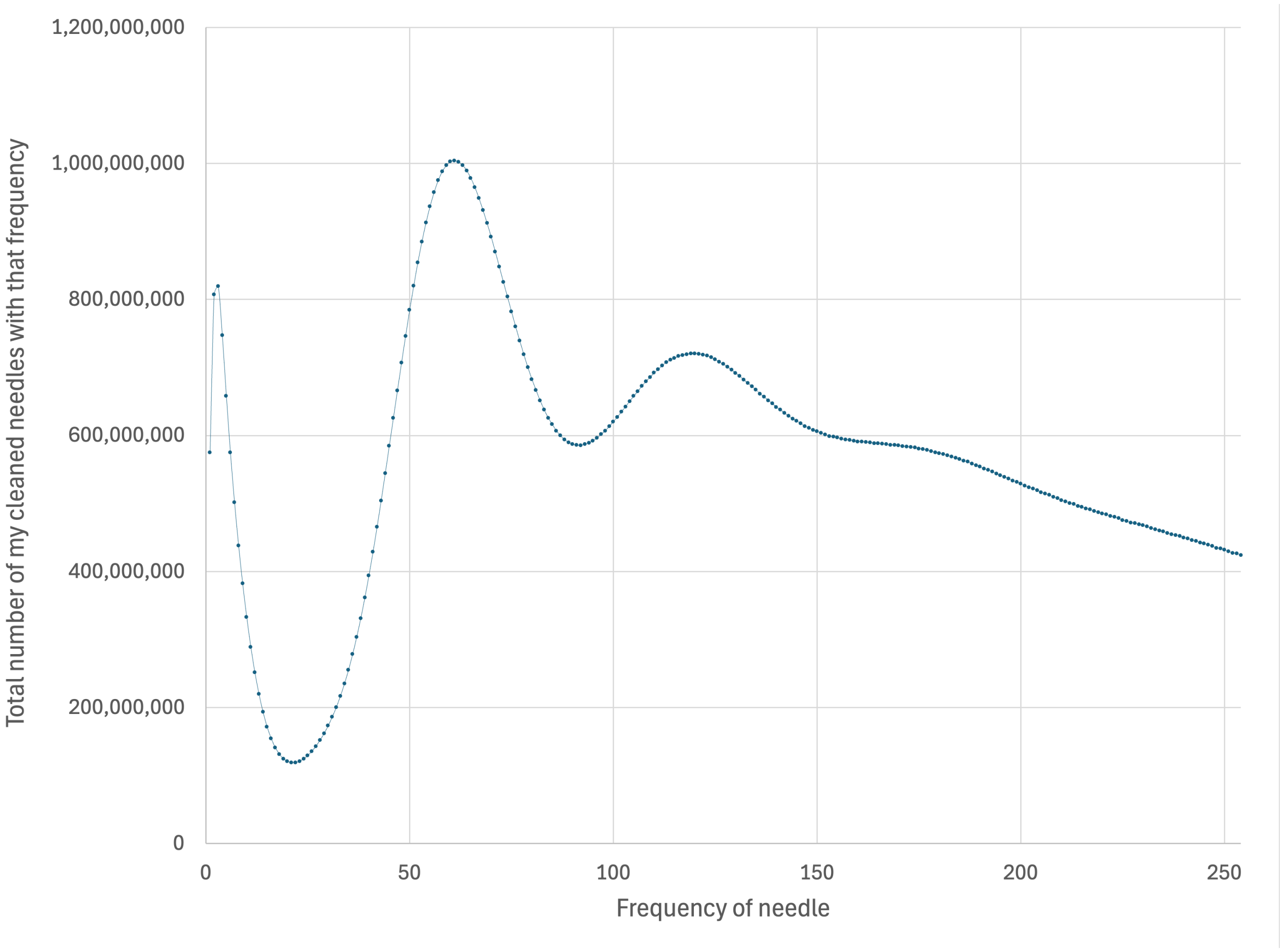

Figure 6: The equivalent of Fig. 1 for my data.

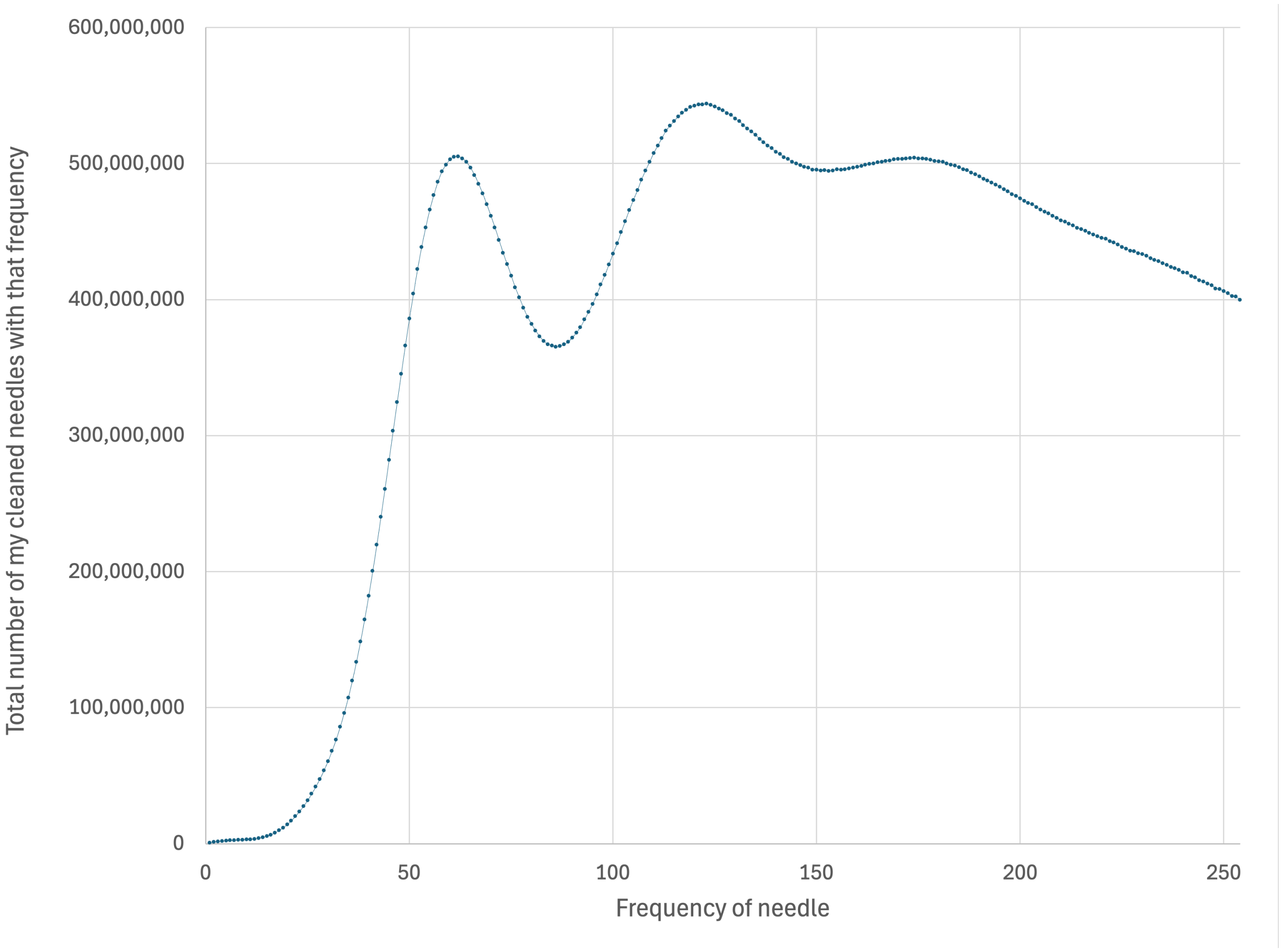

Figure 7: The equivalent of Fig. 2 for my data.

between 1 and 21 inclusive is 7,756,014,090, or 2.18% of my 356,049,894,000 clean needles. Likewise, the equivalent of Fig. 2 is Fig. 7, and the equivalent of Fig. 3 is Fig. 8. In this case the singles peak is at a frequency of 60 for Fig. 7 and 62 for Fig. 8, so I choose a frequency of 61 to split the difference and provide a consistent base for comparison. The number of clean needles with frequencies between 1 and 61 inclusive is then 18,463,287,003 for Fig. 7, and 9,709,635,286 for Fig. 8. The difference 8,753,651,717 is 12.9% larger than the estimate above to the first trough, probably due to both the smaller amount of data (and hence less resolution of the two peaks) and the qualitatively different shape of my curves (I'm assuming due to my Y chromosome). If we add the residual needles between frequencies of 1 and, say, 18 in Fig. 7, namely, 64,754,713, we get an estimate of 8,818,406,430 needles that I consider to be the low-frequency noise. This is 2.48% of my clean needles. Removing them, we have a presumed 347,231,487,570 clean needles that are not just low-frequency noise. Now, the singles peak of Fig. 6 just happens to be at a frequency of 61.00, even fitting a parabola. That gives us an average value of 5.69 billion bases for my two haploid genomes, or 11.38 billion bases for my diploid genome.

The difference between my estimate for Sally of 11.64 billion bases and for me of 11.38 billion bases is about 260 million bases. This is a larger gap than the roughly 100 million that I obtained in the paper. Of course, it is at the limits of precision of my methods, so is subject to change.

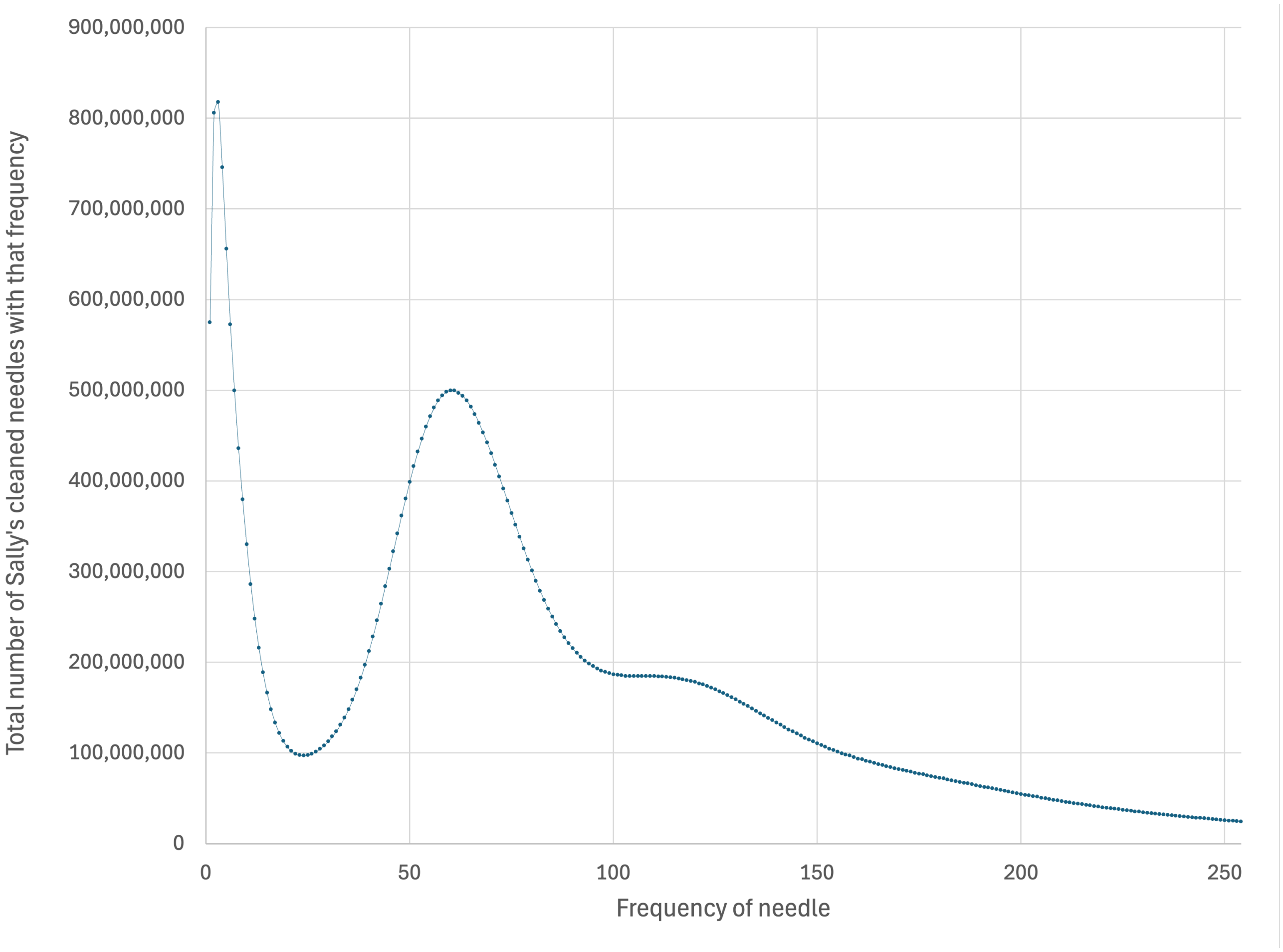

Figure 8: The equivalent of Fig. 3 for my data.

4. Discussion

What I wanted to say in PR1, but was unable to get to, is that any reduction of the diploid genome into the 11-billion-base range is a significant development. As I noted in the paper, current knowledge is that the female diploid genome is 6.37 billion bases long, and the male is 6.27 billion bases long [6]. If Sally's is around 11.64 billion bases, and mine is around 11.38 billion, then there are 5.27 billion missing from Sally's and 5.11 billion missing from mine. Ignoring the bogus precision of these numbers, having *only* 5.2 billion bases missing is big news, at least to me. It means that my original claim that our genomes are twice as long as the reference genome was only approximate; in fact, Sally's is about 83% longer than the reference genome, and mine is about 81% longer. Given the uncertainties, we can split the difference and say that our genomes are about 82% longer than what the NIH is giving us. In other words, about 45% of our genomes are missing in their data.

I neither expected nor wished for this development. The "double, half" story had a nice symmetry to it, even if the reason was mysterious. But when it became apparent that this story was not exactly true, while I was preparing PR1 (even though I have now discarded that early estimate), I realized that it is more heartening than what I originally had. Having *exactly* half the genome missing just seems to point to some sort of data processing or sequencing error. But 45% of it? That *feels* like a number that might just be real.

The flip side, though, is that *almost exactly 50% of singles are missing.* That hasn't changed one iota. I mean, there's no inconsistency there: it just means that the more frequent needles—or some of them, at any rate—are not missing at the same rate, so that *overall* there are about 45% of needles (and hence 45% of bases) missing. But it does beg the question: why is it so very close to 50% for singles?

Obviously, I don't have an answer yet. But I have a gut feeling that the very resonance that got my attention in the first place—that "'Millikan oil drop" set of quanta of Fig. 2 of the paper—is definitely not new. I feel that it has been measured, *somehow*, before—probably many times. And I suspect that the same haploid–diploid confusion that I originally had *just might* be to blame for it being misinterpreted. When I have ventured a guess to Sally or Greg, I liken it to "Who's on First [7]?" And that might just have set in stone the number of bases that researchers expected to find, so that the compiled sequences were fit into a template of that size.

Anyway, that's my guess. We'll just have to see what the answer turns out to be.

References

[1] J. P. Costella, *History of the UnBlur algorithm* (2021).
[2] Google, `doodles.google` (2025).
[3] J. P. Costella, *An amateur analysis of the structure of the human genome* (2025).
[4] J. P. Costella, *The Shadow Genome Project: progress report 1* (2025).
[5] J. P. Costella, `johncostella.com/aliens` (2025).
[6] A. Piovesan *et al.*, *BMC Res. Notes* **12** (2019) 106.
[7] Wikipedia, *Who's on First?* (2025).