The Shadow Genome Project: progress report 1

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

(September 10, 2025)

Abstract

This is my first progress report for The Shadow Genome Project. I report additional analysis that supports my conclusion that the haploid genome is about twice as long as we think it is, but the new method refines the estimate to 11.1 billion bases for Sally's diploid genome, down from the estimate of 12.7 billion in the other paper. I also provide a progress report on building out the first sequences grown from Sally's and my raw data, explaining modifications that I am making to the algorithm outlined in the paper. *(This report is cut short by the breaking news out of Utah.)*

1. Introduction

When I finished the paper [1] on Monday, I thought that I would spend the next three days trying to implement the algorithm that I described in the paper—which, as I described there, I had only started implementing—to try to get some sequences into data files to post to my web page. But my inner skeptic kept nagging at me: are you *really* sure that you haven't fucked it up? I had a nagging feeling that there was another way to double-check the results that might shut that skeptic up, but couldn't quite put my finger on it. (I guess that should be "n-check" at this point. I don't think that there is any upper bound on n.)

Monday and Tuesday night they were telling me that there is a much simpler two-dimensional graph that would at least convince *me* well enough that I could free up my mind to spend all of my spare time on the sequence-growing algorithm, rather than trying to relitigate the paper. When I woke with a start early on Wednesday morning, I had almost all of the details. A bit of trial and error with a new program gave it to me.

But now a quandary: I had an *even better* graph that (with 20:20 hindsight) showed even more simply that Sally's genome is at least 10 billion bases long (uh, more on that in a moment). But the paper was finished; I knew that I wasn't supposed to try to shoehorn any more material into it; it was exactly as it was supposed to be. But I can't just publish it without—as my late friend David Lifton would have said—the "best evidence" in it, right? That's not just burying the lede: it's cremating it and scattering its ashes out on the Potomac.

Then the penny dropped: what's stopping me from publishing *progress reports*? It sounds fucking obvious—right?—but for some reason I hadn't thought of doing for this hobby project what I do routinely at work. I guess I've been brainwashed by the "polish that turd to death" philosophy of academic publishing. But I already broke every rule of academic publication in the paper, so why not just continue on down the same dirt track?

So that's what this is: the first (I assume just the first) progress report. I'll be publishing it tomorrow, Thursday, at the same time as my paper. (And by now you know that by "publishing" I mean just posting it on my website. I mean, it's been 32 years since postdoc Mark Thomson brought me up to his sixth-floor office and excitedly showed me that some kid from the National Center for Supercomputing Applications had written some software so that you could actually *see*, with great ease, those graphs that CERN were making available via that clunky "hypertext transfer protocol" that you had to use command-line tools to retrieve. I didn't realize for decades that I'd recently debated virtual reality with that kid [2]. So I think that it's well and truly time to deem that posting a paper to that new platform is just as valid as getting it typeset by the academic gatekeepers the way Albert Einstein had to.)

2. New analysis

2.1. Frequency bias graphs

Per my comments in Sec. 1, I was led to return to just two-dimensional graphs to provide the cleanest evidence that the doubled-genome hypothesis is not crazy. Ignoring the circuitous path by which I got to the final version, this is the nub of it: Sally's cleaned data has a total of 390,091,247,764 needles in it. The reference genome, done the way I described in Sec. 7 of the paper, has a total of 3,136,840,053. The ratio of these two numbers is about 124.36. (If it weren't for my cleaning process in Sec. 10 of the paper, it would be 421,298,960,365 compared to 3,136,840,053, or 134.31, which is close to Nebula's emailed estimate of 139.05; I'm assuming they either use a different reference genome or a different method of estimating duplicate reads; the difference is of no consequence.)

Now, if you didn't know anything about anything, then you'd assume that any given needle seen in Sally's data should appear about 124 times as often as it does in the reference genome data: that's what this "coverage" of 124 means, on the average. So what if we now actually *compute* this ratio for all 390,091,247,764 needles that appear in Sally's data, and bin the results by the frequency in Sally's data, like Fig. 1 of the paper? We know that an appropriate average of these results (over all frequencies, not just the low ones) must be equal to unity, because of how we've defined it. But *any deviations from unity* will tell us something about the structure of the data. (Oh my god, this is just what Bear does—the idea that let us come to America. Well I'll be buggered.)

Someone just tried to assassinate Charlie Kirk. It's difficult to be excited at all about releasing the papers tomorrow with this on the TV screen near my desk in the LinkedIn office. And takes me back to why I'm releasing it tomorrow [3].

I'll wrap this up quickly.

The program will be included in my codebase tomorrow. The graph of "bias" is shown in Fig. 1. As we saw in Figs. 5 and 6 of the paper, the low-frequency needles are mainly noise—sequencing errors or contaminants—and

Assassinated.

Motherfuckers.

Forget about finishing this. We're done.

So the trough is at 38, the peak is at 81, the next trough is at 109, the next peak at 144, the next trough at 176, the next peak at 212. Dividing by each ordinal, that gives you frequencies of

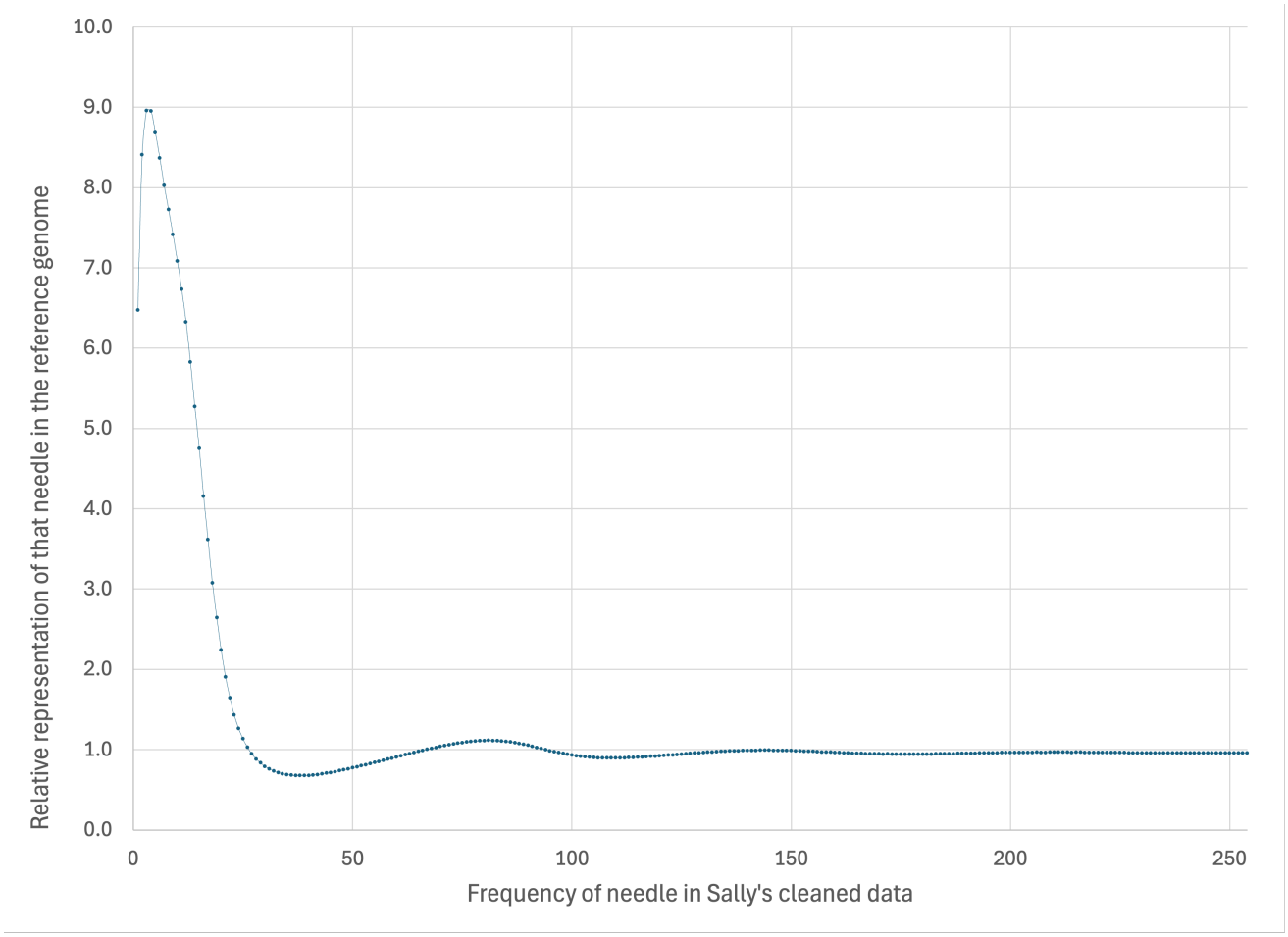

Figure 1: The number of distinct needles for each frequency less than 255 in Sally's data.

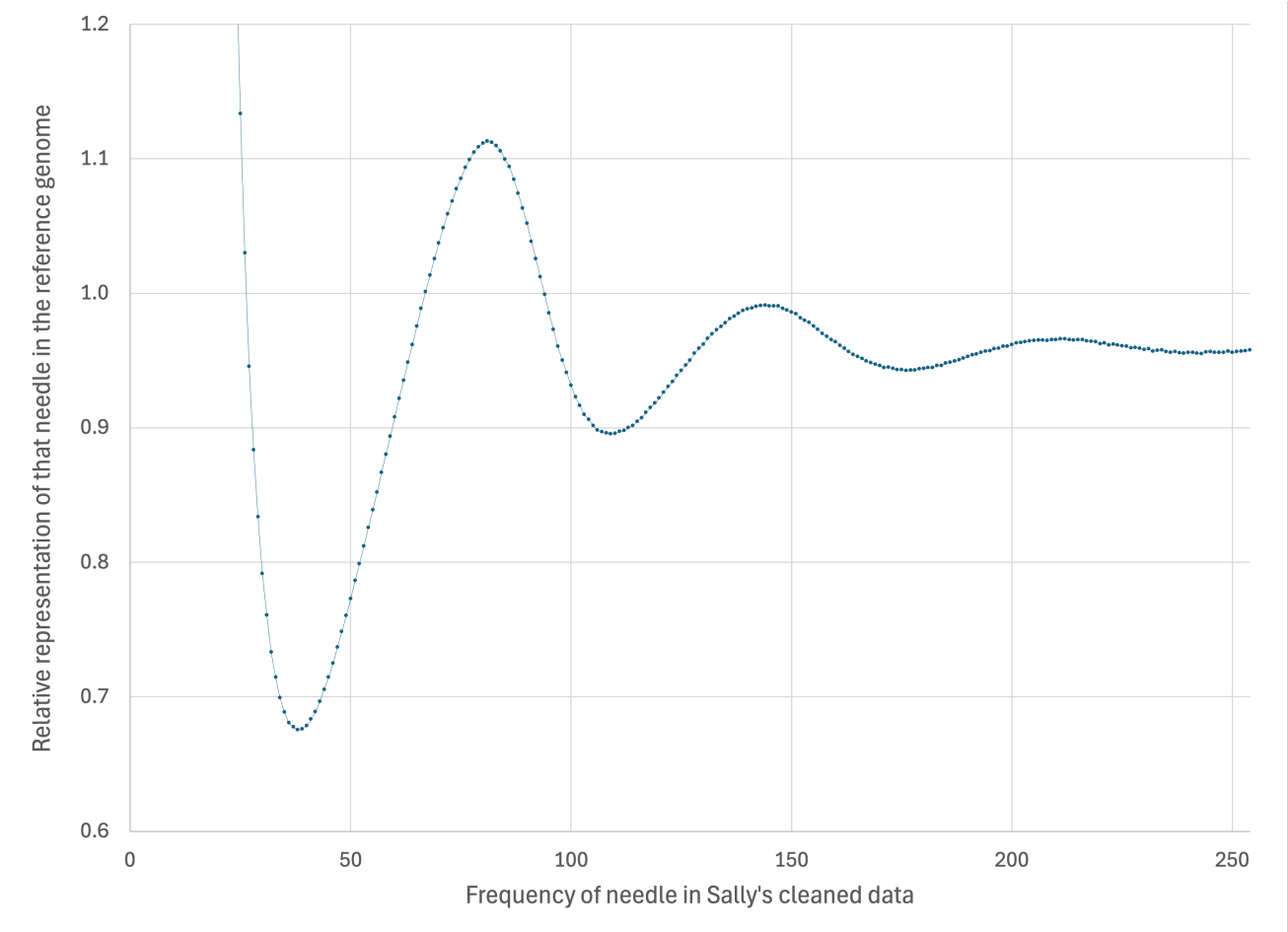

Figure 2: Scaled version of Fig. 1.

38, 40.5, 36.3, 36, 35.2, and 35.3. So about 35.3 out of her 390,091,247,764 needles is 11.1 billion bases.

Better view in Fig. 2.

I think the peaks of the graphs in the other paper—which I equivocated on doing as number of distinct needles or total number of needles—would be pulled up if I did total number of needles, as here. Not sure yet if it will pull it up this high (i.e., dropping the 6.35 billion bases to 5.53 billion).

But who gives a fuck.

3. Sequences grown

(Written before it really sank in.)

I've started growing the sequences. I realized that trying to find the best seed needle by processing all of them is too computationally expensive. Cutting it back, I see sequences over 110 bases long using the method described in the paper.

I will return to this after ... I don't know; I was going to say after we recover from this tragedy, but I don't think we will. Motherfuckers.

This paper is dedicated to Charlie Kirk, assassinated while I was finishing it.

References

[1] J. P. Costella, *An amateur analysis of the structure of the human genome* (2025).
[2] M. L. Andreessen and J. P. Costella, `sci.virtual-worlds` (1992).
[3] J. P. Costella, *History of the UnBlur algorithm* (2021).