The Shadow Genome Project: progress report 4

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

(September 24, 2025)

Abstract

I return to the sequence-growing project. I compute the transition probabilities for the Markov chain of successive unambiguous singles on Sally's genome being on or off the reference genome. The mystery does not subside.

1. Corrections and clarifications

No new corrections or clarifications to report at this time.

2. Markov chains

I noted on page 35 of the paper that my initial attempts at growing seed sequences from unambiguous singles in Sally's data showed that the sequences that started on the reference genome generally stayed on it, and those that were not on it generally stayed off it, but that I had seen occasional transitions from "on-reference" to "off-reference" and vice versa. I gave some examples of these transitions in Figs. 18 and 19.

As I contemplated the task of refactoring the sequence-growing algorithm after we return from our vacation (as I described at the end of PR3), I realized that it would not be difficult to analyze some properties of these transitions *even without growing the actual sequences.* Consider my description in the paper of growing an ambiguous sequence to the right (which in general will cover *all* possible sequence-growing scenarios, if we manage to start with the leftmost needle of each sequence in the first place). At each step we look at the rightmost needle in the sequence grown so far, and then look at the four possible needles obtained by adding A, C, G, or T to the right end of the sequence. If *exactly one* of those four needles is unambiguously the correct one—in the sense that with relatively high confidence we can conclude it to be the next needle, which in the paper I deemed to be the case if the other three needles have frequencies lower than that of the minimum between the noise peak and the singles peaks—then we grow the sequence by appending to it the base in the rightmost position of this new needle. Now, if we *are* able to do that, then we can inspect whether the previous rightmost needle appears in the reference genome, and likewise whether the newly-added needle appears on the reference genome. Indeed, this is how I added the last column in Figs. 16–20 of the paper, for the needle responsible for each base in the grown sequence, including the seed needle.

But there is no reason for us to start with a seed needle. We can walk through *every* possible needle in Sally's data. If it seems to be a single—using, again, say, the same original logic of

1 \ 2	off mine	on mine	total
off mine	7,704,158	5,597,468	13,301,626
on mine	5,117,635	337,079,095	342,196,730

Table 1: Markov chain frequencies for unambiguous singles transitions in Sally's data.

1 \ 2	off mine	on mine	total
off mine	57.919%	42.081%	100%
on mine	1.496%	98.504%	100%

Table 2: The same as Table 1, but normalized into transition probabilities.

having a frequency between the two minima surrounding the singles peak—then we can consider it to be the "left side" needle of a Markov transition. We can then check, using the above logic, whether there is an unambiguous "right side" needle corresponding to it. If so, then we check whether each of the two needles is on the reference genome or not. These two boolean flags allow us to increment the appropriate frequency of this Markov chain model, namely, on-reference to on, on to off, off to on, and off to off. Once we've checked every possible left needle, we can then normalize these frequencies to obtain the corresponding empirical Markov chain transition probabilities.

It sounds simple in retrospect, and it will be almost trivial to program up. Hindsight, eh?

...

OK, I've coded it up. Let's first run it on Sally's clean needles, using *my* clean needles as reference. The results are shown in Table 1. Clearly, the vast majority of the time the singles appearing on Sally's genome are also on mine, and if we start with such a single it stays on my genome most of the time when we transition to the next unambiguous single in Sally's data. We can see that more explicitly by normalizing the frequencies in Table 1 into empirical transition probabilities, which I show in Table 2. We immediately see that expected number of needles that Sally's unambiguous singles will stay on my genome for is $1/1.496\% = 66.87$. This corresponds to about 82 bases. So, for these unambiguous singles sequences in Sally's genome, every 82 bases or so mine will part ways. I assume that these departures are due to variants in our genomes.

It's worth keeping in mind (I have to keep reminding myself to keep this in mind) that we are likely only looking at *special parts* of the genome when we are looking at singles, and even moreso the unambiguous *sequences* of singles. We saw in the paper that Sally's cleaned data has 390,091,247,764 needles in total, with 3,656,593,732 distinct values. The 355,498,356 transitions we're looking at in Table 1 thus represent less than 10% of even the *distinct* needles. Even if we were to make the approximation that their average frequency is the modal value of 61 (I could add up the frequencies explicitly, if needed), that's less than 22 billion total needles out of 390 billion, or less than 6% of them. (OK, I just augmented the program: it tells me that their total frequency is 22,628,598,698, yielding 5.8%.) So Tables 1 and 2 are telling us something about this 5.8% of our genomes. That's a very important 5.8%, I believe, but the other 94.2% could send us a slightly different message.

It's worth keeping this in mind when considering the top row of Table 2. It seems to be telling us that once Sally's sequence of unambiguous sequences falls off mine, then it's almost

2 1	none	off mine	on mine	ambiguous	total
off mine	3,476,232	7,704,158	5,597,468	8,753,350	25,531,208
on mine	115,224,573	5,117,635	337,079,095	264,341,414	721,762,717

Table 3: Expanded Markov chain frequencies for unambiguous singles transitions in Sally's data.

a coin flip for every successive base added as to whether it will return to being on mine again. This seems to make sense, if we think of most variations between our genomes as being single-nucleotide variants. But it *doesn't* actually make sense. Such a variant in one base *should affect 16 consecutive needles.* So how can we flip back almost immediately?

I believe that the answer is that I'm discussing Table 2 as if it governed the entire genome, but it doesn't. It's predicated on singles *that have an unambiguous next single.* That's an assumption that negates my simplistic tale of getting off and back on my genome.

. . .

To try to repair that assumption, I added a different set of Markov chain frequencies to my program. I show them in Table 1.

OK, so four friends just arrived for a surprise birthday party for me. Sally really got me with that one. So I'll cut things there. Luckily for you, I wrote the next section just before that. Enjoy. Cheers!

3. Discussion

As they say in the classics, we're in a right fucking mess. I have no fucking idea what the correct answer is to this mystery.

Over the past fortnight I have expanded slightly my menu of possible explanations. I see at least six options.

One is that Sally and I have alien DNA, unlike that of any other human. I addressed this in Sec. 13 of the paper. Weird as we are, I think it highly unlikely. But we'll let you know after we again return from Roswell.

The second is that I've screwed up something really fundamental. The whole paper was really about that. I think the probability of that is decreasing, but I'll never be so arrogant as to assume that it is anywhere near approaching zero. It remains in the Rumsfeldian domain of unknown unknowns.

The third is that Nebula's machines are just broken. I find that hard to believe. My understanding is that they are just using machines that they source from somewhere; they're not building them from the ground up themselves. They managed to align our raw data onto some version of the reference genome, supplying us with CRAM and CRAI and VCF and TBI files. I have not looked at those files, but they seem to be industry-standard. They also supply online tools for exploring our genetic data. So whatever they are doing, it seems to work. I don't think they could fool too many people for too long if they were doing something wrong, or dodgy.

The fourth is that Nebula's machines are working fine, but their data processing for the two raw FASTQ files is completely broken. This one isn't quite as difficult to believe. Who looks at the raw data anyway? Not many people. But some do. And so it's still difficult to believe. Moreover, Nebula supplied me their own estimates of the coverage of our raw data. It corrected the raw ratio for some things. It would be difficult to understand them doing that if the data files were complete garbage. But I guess it's possible.

The fifth is that Nebula is somehow measuring something more than what people usually measure when they're measuring the human genome. That sounded less eloquent than it did in my head. But you get the idea. They're clearly measuring the right stuff—the alignment to the reference genome is proof of that, as well as the fact that more than half of the needles that they do measure are there in the human genome, somewhere—but maybe they're measuring something more. Or measuring the DNA in a form that is different to what it is for other measurement methods. I can't really see how that would be the case. But I'm no expert. In any case, it's no more crazy than any of the other explanations.

The sixth possibility takes me back to the 2000 National Congress of the Australian Institute of Physics. It was held at the University of Adelaide. Mentone Grammar allowed me to go to it. I drove over. I was sitting in the audience of one of the keynote sessions, held in some hall. Next to me sat Professor Geoff Opat. Largely responsible for me getting to do physics at all, Geoff Opat was the perfect comic-book example of a physics professor. Brilliant. Jewish. Always pushing up his thick glasses. Hair going everywhere. A genius at the real questions of theoretical physics, even though he was actually Professor of Experimental Physics. Anyway, he was sitting beside me. Well, not sitting. He was sleeping. He was fast asleep. He wasn't snoring, but he was close. I was wondering what I'd do if he slowly fell over into my lap. But he didn't.

"Crap!" he exclaimed. The speaker clearly heard it. "That's just crap!" He was wide awake. He'd gone from semi-comatose to mid-sentence before my brain even registered what was happening. He half-leaned half-confidentially my way—it wouldn't have mattered who I was, as long as I was a friend—and started expounding on why what the speaker had said was pure crap. Loud enough that the speaker could hear, but low enough that he was not interrupting the presentation. I can't even remember what the the session was about. Geoff was looking up at the speaker, and his projected presentation, and then back at me, out of the sides of his eyes, pushing up his glasses, and pointing out this and that about what he was saying. There was not an ounce of malice in his words. The speaker was in no way offended. In fact, he looked like he was honored, that the great Geoff Opat had been listening so intently that he had been moved to expostulate so vehemently on some profound truth that he believed the speaker to have violated.

That's why I love physics.

My sixth possible explanation is just like that. A quarter-century later. With the full faith of spirit and charity and honesty of the late Geoff Opat in my veins, I wonder if the reference human genome is just crap. I mean, not in its fine details. We know that we can align our data with it, and figure out the SNPs that each of us have that make us all unique. But I mean in some larger-scale sense. Maybe, somehow, it's just put together wrong. And some of it is missing.

As a number of wise people have said to me—and I wholeheartedly agree with them—it's difficult to see how that could be possible. But just because we can't see it doesn't mean that it's not there. We're not toddlers. I have no concrete vision of how the reference gnome could be so wrong, but I also can't reconcile it with what the data is telling me. So even though some of the other five explanations are probably more likely, I think I will have to proceed, after our vacation, on the assumption that, in Opat's words, it's just crap. And largely ignore it. I feel that I will get far more out of comparing Sally's and my data than I will trying to compare either to the reference genome. That might actually be a benefit, if it turns out that one of the other explanations is correct, because there might still be something to salvage from that relative analysis, even if I have to excise from it whatever errors common to both sets of data it contains.

Anyway, that's what I'll do. It's my party and I'll cry if I want to.