The Shadow Genome Project: progress report 3

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

(September 21, 2025)

Abstract

I shift my attention to the large repeating sequences in the human genome, and show that we can do somewhat more with them than what I had originally believed.

1. Corrections and clarifications

As I noted in PR2, I will include a "Corrections" section in all PRs. I'm now expanding that to include clarifications as well.

1.1. Bases, base pairs, haploid genomes, and diploid genomes

As my good friend David Mantik has pointed out, readers of my papers may be confused by my use of the term "bases" rather than "base pairs," and might mistakenly believe that I've just confused one for the other. I haven't, but I probably haven't made that clear. So now I will. Or I'll try to, at least.

There seems to be a horribly confusing set of inconsistencies in how practitioners in this field refer to the nucleotides. Technically, `A`, `C`, `G`, and `T` are the "bases," but the actual physical double-helix is of course made up of a "twisted ladder" of base *pairs*, where the "complementary" strand always has the complementary base (`A` with `T`, and `C` with `G`). But this means that the physical double helix contains *no more information* than each of its two constituent strands. Now, there are many areas of research where the physical structure of the double helix—and particularly its chemistry—are of interest, or indeed are even the sole focus. But for *bioinformatics* purposes there is no point in explicitly describing or storing the list of bases in both of the complementary strands, as they contain exactly the same information, just in "mirror image" form. So there is a standard convention for reading out the sequence from one of the ends of one of the strands. The complementary strand can be inferred from it.

The confusion comes from some schools of research insisting on referring to, say, the bioinformatical `A` as a "base pair," whereas the other dissenting schools of research just refer to it as a "base." Everyone generally understands what everyone is talking about, but to a beginner (like me!) there is plenty of scope for confusion.

To make things absolutely clear, the terminology that I have absorbed from the interwebs, and have standardized myself on, is as follows: We each inherit one haploid genome from our mother, and one from our father. Each haploid genome is made up of 23 disconnected chromosomes. Our diploid genome consists of the two haploid genomes that we inherit from our parents. It has

46 chromosomes (except for chromosomally anomalous people). The reference human haploid genome has about 3.15 billion bases. This means that its physical structure is made up of about 3.15 billion base pairs, or about 6.3 billion nucleotide molecules laddered in a double helix. The reference human diploid genome therefore has about 6.3 billion bases, so that its physical structure is made up of about 6.3 billion base pairs, or about 12.6 billion nucleotide molecules.

I found, in my original paper, that Sally's and my genomes are about twice as long as the reference genome. I refined this in PR2 to slightly less than twice as long—namely, about 1.82 times as long. We each have about 5.75 billion bases on each of our haploid genomes. Each of them is physically made up of about 5.75 billion base pairs, or 11.5 billion nucleotides laddered in a double helix. Our diploid genome is therefore about 11.5 billion bases long, so that its physical double-helix structure has about 23 billion nucleotide molecules in total.

The exact numbers here are not meant to imply this level of precision. What is important is that it doesn't matter which way you look at it—23 compared to 12.6, or 11.5 compared to 6.3, or 5.75 compared to 3.15—what is important is that I inferred that Sally's and my genomes have 82% more bases in them than what the NIH has given to us. I might be wrong in that, for many reasons, but *not* because I have confused bases for base pairs.

I will continue to use only the term "bases" throughout these PRs, and will not mention "base pairs" again. For my sanity, at least, if not yours.

2. The most popular needles

Sally couldn't believe her eyes when she turned on the TV today and saw both of our Heads of State traveling together in a horse-drawn carriage to Windsor Castle. But their combined popularity—even added to that of our future King and Queen, who had welcomed our President and First Lady at the airfield—is nothing compared to how popular some needles are in the human genome.

OK, my humor engine is not really firing on all cylinders yet. At least I'm trying.

From the moment I turned my attention to using hexadecigram needles on the genome bases themselves on August 17, I knew that they would be essentially useless for the large repeating sections of the genome. It's simply not feasible to infer how to piece together such large sequences from 16-base samples: there is no way to figure out how many repeats there should be before we "get to the other end."

So I haven't spent much time thinking about them. But after realizing on Tuesday that the diploid genome is only 11.5 billion bases long—not the 12.7 billion bases I originally estimated—I started thinking more about those large, useless, repeating sequences—essentially, the bulk fiber or roughage in our genome. (Sorry, I had Sultana Bran for breakfast this morning—or "Raisin Bran" as the locals like to call it; apparently, sultanas were deleted from the local vernacular at some point in time.)

The first thing I wanted to know was: just *which* needles are the most popular in our genomes? It's not a difficult program to write. The 40 most popular in Sally's are listed in Table 1, and the top 40 for me are shown in Table 2. Note that I am now color-coding the bases, in a way that even my genetically variant eyes can see. If you flip between the two tables, you'll see that the list is pretty consistent between the two of us, with some clearly-related groups of needles permuting positions within their respective groups, and sometimes groups with similar frequencies mingling with each other.

needle	frequency
AAAAAAAAAAAAAAAAAAAA	153,521,525
TTTTTTTTTTTTTTTTTTTT	126,734,038
TCCATTCCATTCCATT	50,349,215
TTCCATTCCATTCCAT	49,056,541
ATTCCATTCCATTCCA	47,589,217
CCATTCCATTCCATTC	45,951,415
CATTCCATTCCATTCC	45,494,474
AATGGAATGGAATGGA	40,792,156
ACACACACACACACACACAC	40,043,808
ATGGAATGGAATGGAA	39,918,441
CACACACACACACACACACA	39,528,715
TGGAATGGAATGGAAT	38,085,080
GAATGGAATGGAATGG	37,000,902
GGAATGGAATGGAATG	36,930,545
GTGTGTGTGTGTGTGTGTGT	30,854,193
TGTGTGTGTGTGTGTGTGTG	30,541,459
TTTTGTATTTTTAGTA	27,486,223
CCCCCCCCCCCCCCCCCCCC	27,068,571
TTTGTATTTTTAGTAG	26,429,711
TGCACTCCAGCCTGGG	26,368,307
TTGTATTTTTAGTAGA	26,021,053
TGTATTTTTAGTAGAG	25,214,279
TTTTTGTATTTTTAGT	24,966,129
CAAAGTGCTGGGATTA	24,863,038
TAATCCCAGCACTTTG	24,741,902
GCCTCCCAAAGTGCTG	24,682,035
GTATTTTTAGTAGAGA	24,557,032
CACTGCACTCCAGCCT	24,427,756
CCAAAGTGCTGGGATT	24,328,213
AATCCCAGCACTTTGG	24,171,296
TCCCAAAGTGCTGGGA	24,059,172
CCTCCCAAAGTGCTGG	23,975,189
CAAAAAAAAAAAAAAAAAAA	23,853,237
CCCAAAGTGCTGGGAT	23,806,208
ACTGCACTCCAGCCTG	23,796,763
CAGCACTTTGGGAGGC	23,776,906
TACTAAAAAATACAAAA	23,693,601
ATCCCAGCACTTTGGG	23,687,243
AAGTGCTGGGATTACA	23,519,576
CTCCCAAAGTGCTGGG	23,517,290

Table 1: The most popular 40 needles in Sally's cleaned data.

needle	frequency
AAAAAAAAAAAAAAAAAAAA	143,753,524
TTTTTTTTTTTTTTTTTTTT	114,073,919
TCCATTCCATTCCATT	44,143,952
TTCCATTCCATTCCAT	43,064,219
ATTCCATTCCATTCCA	41,414,884
CCATTCCATTCCATTC	39,795,884
CATTCCATTCCATTCC	39,332,117
ACACACACACACACACACAC	35,683,450
CACACACACACACACACACA	35,222,972
AATGGAATGGAATGGA	34,800,161
ATGGAATGGAATGGAA	34,126,636
TGGAATGGAATGGAAT	32,214,188
GAATGGAATGGAATGG	31,145,561
GGAATGGAATGGAATG	31,064,214
GTGTGTGTGTGTGTGT	27,251,163
TGTGTGTGTGTGTGTG	26,970,901
TTTTTGTATTTTTTAGTA	24,301,807
TGCACTCCCAGCCCTGGG	24,084,680
TTTGTATTTTTTAGTAG	23,347,839
TTGTATTTTTTAGTAGA	22,983,373
CAAAGTGCTGGGGATTA	22,627,096
CACTGCACTCCCAGCCT	22,476,995
GCCTCCCCAAAGTGCTG	22,341,478
TGTATTTTTTTAGTAGAG	22,223,986
CCAAAGTGCTGGGGATT	22,130,067
TTTTTGTATTTTTTAGT	22,062,713
CAAAAAAAAAAAAAAAAAAA	22,026,232
ACTGCACTCCCAGCCTG	21,894,246
TCCCAAAGTGCTGGGGA	21,815,732
CCCCCCCCCCCCCCCCCCCC	21,765,523
CCTCCCCAAAGTGCTGG	21,672,514
GTATTTTTTTAGTAGAGA	21,625,794
TACTAAAAAATACAAAAA	21,620,319
CCCAAAGTGCTGGGGAT	21,604,200
CCACTGCACTCCCAGCC	21,603,289
TAATCCCCAGCACTTTG	21,596,562
AAGTGCTGGGGATTACA	21,415,347
CTCCCCAAAGTGCTGGG	21,271,434
CTGCACTCCCAGCCTGG	21,238,961
AAAGTGCTGGGGATTAC	21,154,046

Table 2: The most popular 40 needles in my cleaned data.

Now, I knew from my very first work on August 17 that we each have at least one needle that appears more than a hundred million times in our data. But when I first saw the data for Table 1, and realized that that needle is A A A A A A A A A A A A A A A A—which in my encoding is represented by the integer zero—I immediately Scooby Doo-ed, "Ruh Roh." Could that most popular needle be due to some sort of bug in one of my programs, that somehow defaulted to a needle value of zero when the bug was triggered? I racked my brain but couldn't think of any obvious way that that could happen, even with the most careless of bugs, without totally destroying how the "needle photocopier" worked. But I was still suspicious enough that I went quacking (I used to say "googling," but that isn't really accurate anymore since I only use DuckDuckGo; consulted for an opinion, Sally immediately exclaimed, "Quacking!" and her decision is final) for the number of times that sequences of consecutive A bases appear in the reference genome. Not finding a satisfactory answer in a small enough number of seconds that even a Gen α would be proud of my impatience (and confirming that googling would not have been any better than quacking) I decided that computing it myself from the reference genome was probably the answer that the Universe wanted me to arrive at in the first place.

So I added an *ad hoc* lookup of that to the program, intending to remove it again once I had the number. I was heartened to find that A A A A A A A A A A A A A A A A appeared 1,059,352 times in the reference genome. A quick calculation confirmed that receipt of heart: its frequency in Sally's data is about 145 times that in the reference genome, and in mine about 136. The exact numbers don't mean anything much, but they are definitely in the right ballpark.

Of course, I was using the same collection class in the C programs for the reference genome's frequencies as for Sally's and mine, so it was possible that there could be a bug there; but since the reference genome's frequencies had been computed in a totally different way in a totally different program (as described on pages 7 and 8 of the paper, to handle ambiguous bases), I thought that such a bug would be unlikely, given the simplicity of the frequencies collection class (just a wrapper around the raw 16 GiB C array). But, regardless, as another sanity check I went back to the `grep` confirmation method that I described on pages 40 and 41 of the paper:

```
grep -i AAAAAAAAAAAAAAAA GCF_000001405.40_GRCh38.p14_genomic.fna | wc -l
```

which tells me that a run of at least 16 A s appears on 161,670 of the 80-base lines of the reference FASTA file. It's not possible from just that one command to know how many of those represent a run of exactly 16 A s, how many represent a run of 17, how many represent multiple such runs on the same line, *etc.*, all the way up to all 80 of the bases on that line being A s. But with `grep` we *can* actually tease that out a little: `grep`ping for *17* A s yields just 132,397 lines, so there are only 29,273 lines that contain one or more runs of exactly 16 A s. And so on: 23,449 lines with runs of 18, *etc.* I unthinkingly went to write a program to automate all this, until I remembered that I was using `grep` to be *independent* of all of my dodgy programs; so I continued on, old school, running `grep` manually in multiple Terminal windows. I ultimately proved to myself that this analysis establishes that there are at least 858,981 A A A A A A A A A A A A A A A A needles in the reference genome, using nothing more than the command line. That's more than consistent with my program's lookup result of 1,059,352 such needles. I was happy.

This manual analysis also surprised me: I had expected to see long runs of A s, but the ones in the reference genome are relatively short: about half of the needles are in runs less than 23 A s long; there are only three lines with a run of at least 59; and just two with a run of at least 62, both of which turned out to be runs of length 68. So, being paranoid has had the by-product

of teaching me something else about the reference genome (at the very least)—something that would be impossible for me to determine from just the 16-base needles.

All good. So having convinced myself that this most popular needle is consistent with the reference genome, I then realized that the ratio of 145 that I had calculated—of its frequency in Sally's data compared to that in the reference genome—would be useful to see for *all* of the 40 popular needles. I expected that this ratio would be roughly constant for all of them. My argument to myself was that these roughage needles should appear in approximately the same ratio as the total number of (clean) needles in Sally's data compared to the number of needles in the reference genome—namely, 390,091,247,764 to 3,136,840,053, or about 124. (Note that this "coverage" ratio, *compared to the reference genome*, is independent of how long we find the haploid genome to *actually* be; it corresponds to Nebula's quoted coverage of about 139, but remember that I've cleaned off the split ends, and there was always a small residual discrepancy with how Nebula estimated their ratios.) Against that expectation, 144 seemed a little high, but I was willing to wave that away in my mind as something that probably had an easy explanation.

Anyway, I simply added the lookup of the reference frequency and the calculation of the ratio to the main part of my program, and output them in new columns in the tables. I show the results in Tables 3 and 4.

Well.

Well, that's a fucking surprise.

I've discussed the first row—the runs of A s—above. There's then a row for runs of T s, with a ratio of about 119 (all ratios here are for Sally). OK, closer to my expected 124. But then there are five needles that clearly represent repeating C C A T T sequences (Sally miaowed loudly at those—my pet name for her is "pussycat"; hers for me is "puppydog," but I don't have any representation here except maybe the G), but *each with ratios between 3,100 and 3,500.* What the fuck? Then there are another five rows for G G A A T (almost her pet name in Spanish? but I might be pushing my luck there) with ratios between 1,900 and 2,300. Again, what the fuck? Now, it's somewhat confusing at first glance, but about as popular in her data—and hence mixed in with the needles for G G A A T—are two needles for A C, with ratios of 114.6 and 114.7. (These two groups are actually slightly separated in my data, so are easier to visually untangle in Table 2 than in Table 1.) So now we're back to ratios that make more sense. We then have two for G T, with ratios of 87.4 and 87.7. OK, a little low, but still sane. Skipping the next one, for the moment, we see a row for runs of C s—*with a ratio of about 33,000.* What the *fucking* fuck?

Clearly my assumption of a relatively constant ratio of 124 was . . . how can I say it? *Fucked.* If there is nothing radically wrong with my programs, then Tables 1 and 2 are probably telling us something profound about the shadow genome. Its roughage—or at least this end of it—seems to have a much different constitution than its counterpart in the reference genome. To be sure, the total number of needles in Sally's Top 40 of Table 1 or 3 is only 1,445,402,454, or 0.37% of her data. But still, to have such wild differences in prevalence is shocking. At least it is to me.

Anyway, let's swallow our shock, and return to Table 3. The needle I skipped above is clearly part of some repeating sequence that is longer than the 16 bases we see in these needles, since we see shifted versions of it in some of the needles further down. But note that the frequencies are beginning to get very dense, so the remaining counterparts of these needles almost certainly appear slightly further down the table than just the top 40 rows I'm showing here. There are other sets of needles that are clearly related, but the number of different sequences shown in

needle	frequency	reference	ratio
AAAAAAAAAAAAAAAAAAAA	153,521,525	1,059,352	144.9
TTTTTTTTTTTTTTTTTTTT	126,734,038	1,067,748	118.7
TCCATTCCATTCCATT	50,349,215	16,027	3141.5
TTCCATTCCATTCCAT	49,056,541	14,234	3446.4
ATTCCATTCCATTCCA	47,589,217	14,703	3236.7
CCATTCCATTCCATTC	45,951,415	14,624	3142.2
CATTCCATTCCATTCC	45,494,474	13,646	3333.9
AATGGAATGGAATGGA	40,792,156	20,615	1978.8
ACACACACACACACACACAC	40,043,808	349,486	114.6
ATGGAATGGAATGGAA	39,918,441	17,022	2345.1
CACACACACACACACACACA	39,528,715	344,684	114.7
TGGAATGGAATGGAAT	38,085,080	17,297	2201.8
GAATGGAATGGAATGG	37,000,902	18,137	2040.1
GGAATGGAATGGAATG	36,930,545	16,188	2281.4
GTGTGTGTGTGTGTGTGTGT	30,854,193	353,042	87.4
TGTGTGTGTGTGTGTGTGTG	30,541,459	348,144	87.7
TTTTGTATTTTTAGTA	27,486,223	195,106	140.9
CCCCCCCCCCCCCCCCCCCC	27,068,571	828	32691.5
TTTGTATTTTTAGTAG	26,429,711	188,411	140.3
TGCACTCCAGCCTGGG	26,368,307	195,024	135.2
TTGTATTTTTAGTAGA	26,021,053	187,191	139.0
TGTATTTTTAGTAGAG	25,214,279	180,525	139.7
TTTTTGTATTTTTAGT	24,966,129	176,577	141.4
CAAAGTGCTGGGATTA	24,863,038	220,844	112.6
TAATCCCAGCACTTTG	24,741,902	220,254	112.3
GCCTCCCAAAGTGCTG	24,682,035	215,512	114.5
GTATTTTTAGTAGAGA	24,557,032	176,935	138.8
CACTGCACTCCAGCCT	24,427,756	179,362	136.2
CCAAAGTGCTGGGATT	24,328,213	216,346	112.5
AATCCCAGCACTTTGG	24,171,296	215,847	112.0
TCCCAAAGTGCTGGGA	24,059,172	214,093	112.4
CCTCCCAAAGTGCTGG	23,975,189	209,990	114.2
CAAAAAAAAAAAAAAAAAAA	23,853,237	159,326	149.7
CCCAAAGTGCTGGGAT	23,806,208	212,634	112.0
ACTGCACTCCAGCCTG	23,796,763	175,460	135.6
CAGCACTTTGGGAGGC	23,776,906	214,809	110.7
TACTAAAAAATACAAAA	23,693,601	194,805	121.6
ATCCCAGCACTTTGGG	23,687,243	211,788	111.8
AAGTGCTGGGATTACA	23,519,576	210,978	111.5
CTCCCAAAGTGCTGGG	23,517,290	207,144	113.5

Table 3: The most popular 40 needles in Sally's cleaned data, compared to the reference genome.

needle	frequency	reference	ratio
AAAAAAAAAAAAAAAAAAAA	143,753,524	1,059,352	135.7
TTTTTTTTTTTTTTTTTTTT	114,073,919	1,067,748	106.8
TCCATTCCATTCCATT	44,143,952	16,027	2754.3
TTCCATTCCATTCCAT	43,064,219	14,234	3025.4
ATTCCATTCCATTCCA	41,414,884	14,703	2816.8
CCATTCCATTCCATTC	39,795,884	14,624	2721.3
CATTCCATTCCATTCC	39,332,117	13,646	2882.3
ACACACACACACACACACAC	35,683,450	349,486	102.1
CACACACACACACACACACA	35,222,972	344,684	102.2
AATGGAATGGAATGGA	34,800,161	20,615	1688.1
ATGGAATGGAATGGAA	34,126,636	17,022	2004.9
TGGAATGGAATGGAAT	32,214,188	17,297	1862.4
GAATGGAATGGAATGG	31,145,561	18,137	1717.2
GGAATGGAATGGAATG	31,064,214	16,188	1919.0
GTGTGTGTGTGTGTGTGTGT	27,251,163	353,042	77.2
TGTGTGTGTGTGTGTGTGTG	26,970,901	348,144	77.5
TTTTGTATTTTTTAGTA	24,301,807	195,106	124.6
TGCACTCCAGCCTGGG	24,084,680	195,024	123.5
TTTGTATTTTTTAGTAG	23,347,839	188,411	123.9
TTGTATTTTTTAGTAGA	22,983,373	187,191	122.8
CAAAGTGCTGGGATTA	22,627,096	220,844	102.5
CACTGCACTCCAGCCT	22,476,995	179,362	125.3
GCCTCCCAAAGTGCTG	22,341,478	215,512	103.7
TGTATTTTTTAGTAGAG	22,223,986	180,525	123.1
CCAAAGTGCTGGGATT	22,130,067	216,346	102.3
TTTTTGTATTTTTTAGT	22,062,713	176,577	124.9
CAAAAAAAAAAAAAAAAAAA	22,026,232	159,326	138.2
ACTGCACTCCAGCCTG	21,894,246	175,460	124.8
TCCCAAAGTGCTGGGA	21,815,732	214,093	101.9
CCCCCCCCCCCCCCCCCCCC	21,765,523	828	26286.9
CCTCCCAAAGTGCTGG	21,672,514	209,990	103.2
GTATTTTTTAGTAGAGA	21,625,794	176,935	122.2
TACTAAAAAATACAAAA	21,620,319	194,805	111.0
CCCAAAGTGCTGGGAT	21,604,200	212,634	101.6
CCACTGCACTCCAGCC	21,603,289	172,818	125.0
TAATCCCAGCACTTTG	21,596,562	220,254	98.1
AAGTGCTGGGATTACA	21,415,347	210,978	101.5
CTCCCAAAGTGCTGGG	21,271,434	207,144	102.7
CTGCACTCCAGCCTGG	21,238,961	171,322	124.0
AAAGTGCTGGGATTAC	21,154,046	207,266	102.1

Table 4: The most popular 40 needles in my cleaned data, compared to the reference genome.

those 40 rows is clearly proliferating.

The one sequence that I was hoping to compute later (I haven't told you that bit yet)—and didn't at all expect to see in this Top 40—is C A A A A A A A A A A A A A A A A A. I figured that, no matter how long the repeating A sequences happen to be, *they have to end somewhere*. By looking up the frequencies of those needles that straddle one base outside of the sequence of As (here a C on the left), and maybe double-checking with the next base as well, I hoped to be able to at least estimate *how many* such repeating sequences there are in Sally's haploid genome, even if I can't figure out *how long* any of them is. In other words, if we think of each such run of As terminated on each end by non-As as a jump rope, then I can measure—just from the needle frequencies—how many *handles* there are at the ends of the jump ropes, in total. The number of handles is of course just twice the number of jump ropes. We already have the total length of rope needed to make the set of jump ropes, but we won't be able to figure out (from just the needle frequencies) the length of any *given* jump rope, other than that it must be at least 16 As long. All we can compute is the *mean* length of those jump ropes, which I had thought would be a useful statistic.

So, still ignoring my shock, and plowing ahead with that original goal, I just wrote a quick program to look up the frequencies of all the relevant needles, at least for those jump ropes of a solid color—henceforth, "solid ropes." The results for Sally are shown in Table 5, and those for me in Table 6.

Clearly, for runs of As and Ts, the ratios for handles make sense, are roughly in agreement with those for the ropes, and aren't far off the overall ratio of 124 for Sally's data. It seems like they are not too badly represented in the reference genome. For runs of Gs, the picture is less clear, with ratios having a much wider variation, with values in the hundreds or thousands; clearly, these are not as well handled in the reference genome. For runs of Cs, we again see ratios of thousands to the tens of thousands; the reference genome seems to be a basket case. Of course, there may well be position-dependent biases that affect the sampling of Nebula's machine for these runs of bases. But it is difficult to see how that could make such a huge difference.

Let's table those observations for now; I have to think about it some more. Returning to the question of how many handles there are for the jump ropes, we see that there are 30,608,451 C, G, and T left handles for runs of As, and 30,375,219 right handles. The latter is 99.2% of the former; I have to assume that the remaining 0.8% represent the balance of runs of As that reach the right end of a (normally 125-needle) sequence compared to those that reach the left end. Having about 30 million of each handle for 153 needles of rope tells us that those jump ropes have a mean of about 5.0 needles, which means a run of 20 consecutive As. For the reference genome we have 222,254 left handles and the same number of right handles (which makes sense, because the sequences in the referencce genome are hugely longer than they are for the raw Nebula data). For 1,059,352 needles of rope that implies a mean number of needles in the jump rope of 4.77, or a mean run of about 19.8 As. This is consistent with my direct `grep` analysis of the reference genome, from which I derive a possible interval of $[19.47, 24.01]$ for this mean. It's also very close to the mean of 20 that I find in Sally's data. I conclude that there is no real discrepancy for runs of As between Sally's data and the reference genome.

Likewise, for runs of Ts, we see 28,227,112 A, C, and G left handles and 28,148,621 right handles; the latter is 99.7% of the former, which again is reasonable for these short sequences. For 127 million needles of rope, that gives us a mean run of Ts of about 4.49 needles, or about 19.5 bases. The reference genome has 223,724 left handles and 223,727 right handles—a difference of

needle	frequency	reference	ratio
AAAAAAAAAAAAAAAAAAAA	153,521,525	1,059,352	144.9
CAAAAAAAAAAAAAAAAAAA	23,853,237	159,326	149.7
GAAAAAAAAAAAAAAAAAAA	3,253,627	20,882	155.8
TAAAAAAAAAAAAAAAAAAA	6,085,186	42,046	144.7
AAAAAAAAAAAAAAAAAAAC	4,291,603	28,390	151.2
AAAAAAAAAAAAAAAAAAAG	20,332,407	138,258	147.1
AAAAAAAAAAAAAAAAAAAT	8,394,305	55,606	151.0
CCCCCCCCCCCCCCCCCCCC	27,068,571	828	32691.5
ACCCCCCCCCCCCCCCCCCC	997,661	119	8383.7
GCCCCCCCCCCCCCCCCCCC	1,366,091	199	6864.8
TCCCCCCCCCCCCCCCCCCC	1,471,921	92	15999.1
CCCCCCCCCCCCCCCCCCCA	1,129,565	159	7104.2
CCCCCCCCCCCCCCCCCCCG	1,339,805	239	5605.9
CCCCCCCCCCCCCCCCCCCT	1,069,335	12	89111.2
GGGGGGGGGGGGGGGGGGGG	558,762	957	583.9
AGGGGGGGGGGGGGGGGGGG	35,153	22	1597.9
CGGGGGGGGGGGGGGGGGGG	86,030	243	354.0
TGGGGGGGGGGGGGGGGGGG	56,466	126	448.1
GGGGGGGGGGGGGGGGGGGA	48,486	87	557.3
GGGGGGGGGGGGGGGGGGGC	83,273	161	517.2
GGGGGGGGGGGGGGGGGGGT	36,201	148	244.6
TTTTTTTTTTTTTTTTTTTT	126,734,038	1,067,748	118.7
ATTTTTTTTTTTTTTTTTTT	7,523,931	56,337	133.6
CTTTTTTTTTTTTTTTTTTT	17,063,780	138,774	123.0
GTTTTTTTTTTTTTTTTTTT	3,639,401	28,613	127.2
TTTTTTTTTTTTTTTTTTTA	5,510,232	41,997	131.2
TTTTTTTTTTTTTTTTTTTC	2,755,352	20,918	131.7
TTTTTTTTTTTTTTTTTTTG	19,883,037	160,812	123.6

Table 5: The relevant needles for the jump ropes of a single solid color in Sally's cleaned data.

needle	frequency	reference	ratio
AAAAAAAAAAAAAAAAAAAA	143,753,524	1,059,352	135.7
CAAAAAAAAAAAAAAAAAAA	22,026,232	159,326	138.2
GAAAAAAAAAAAAAAAAAAA	2,976,465	20,882	142.5
TAAAAAAAAAAAAAAAAAAA	5,605,754	42,046	133.3
AAAAAAAAAAAAAAAAAAAC	3,952,757	28,390	139.2
AAAAAAAAAAAAAAAAAAAG	18,725,028	138,258	135.4
AAAAAAAAAAAAAAAAAAAT	7,697,434	55,606	138.4
CCCCCCCCCCCCCCCCCCCC	21,765,523	828	26286.9
ACCCCCCCCCCCCCCCCCCC	818,823	119	6880.9
GCCCCCCCCCCCCCCCCCCC	1,086,689	199	5460.7
TCCCCCCCCCCCCCCCCCCC	1,230,423	92	13374.2
CCCCCCCCCCCCCCCCCCCA	944,064	159	5937.5
CCCCCCCCCCCCCCCCCCCG	1,065,841	239	4459.6
CCCCCCCCCCCCCCCCCCCT	895,296	12	74608.0
GGGGGGGGGGGGGGGGGGGG	409,302	957	427.7
AGGGGGGGGGGGGGGGGGGG	26,966	22	1225.7
CGGGGGGGGGGGGGGGGGGG	60,963	243	250.9
TGGGGGGGGGGGGGGGGGGG	43,897	126	348.4
GGGGGGGGGGGGGGGGGGGA	36,230	87	416.4
GGGGGGGGGGGGGGGGGGGC	61,661	161	383.0
GGGGGGGGGGGGGGGGGGGT	27,631	148	186.7
TTTTTTTTTTTTTTTTTTTT	114,073,919	1,067,748	106.8
ATTTTTTTTTTTTTTTTTTT	6,693,996	56,337	118.8
CTTTTTTTTTTTTTTTTTTT	15,186,522	138,774	109.4
GTTTTTTTTTTTTTTTTTTT	3,243,226	28,613	113.3
TTTTTTTTTTTTTTTTTTTA	4,958,199	41,997	118.1
TTTTTTTTTTTTTTTTTTTC	2,446,965	20,918	117.0
TTTTTTTTTTTTTTTTTTTG	17,606,656	160,812	109.5

Table 6: The relevant needles for the jump ropes of a single solid color in my cleaned data.

3 that must again represent running up against a sequence break—and compared to the 223,727 rope needles gives us a mean number of needles per jump rope of 4.77, or about 19.8 bases. Again, we get good consistency between Sally and the reference genome.

For runs of G s, we have 177,649 left handles in Sally's data, and 167,960 right handles. The latter is about 95% of the former, which is a wider variation than I would have expected to see. Taking the larger, former number with the 558,762 rope needles gives me a mean of 3.1 needles per run, or 18.1 bases. The reference genome has 391 left handles and 396 right handles, which for 957 rope needles gives a mean of 2.4 needles, or 17.4 bases. The difference in the means may be significant, given how wrong the ratios are, but it's difficult to be definitive.

Even more crazy, of course, are the runs of C s. In Sally's data there are 3,835,673 left handles and 3,538,705 right handles, which is now nearly an 8% difference. Again taking the former number with the 27,068,571 rope needles gives a mean number of needles per jump rope of about 7.1, *i.e.*, a mean run of 22.1 bases. For the reference genome there are 410 left handles and the same number of right handles, which with the 828 rope needles gives an average of 2.0 needles per jump rope, or 19 bases. Again, I can't be sure if this difference in mean is significant, because the ratio is so far off that we're definitely looking only a small subset of them for the reference genome.

I just realized that I made an assumption, early on in this method, that these solid-color runs would be long. Clearly, they are not; they are relatively short. A consequence of this assumption is that the handle needles in Tables 5 and 6, which may actually represent runs only *15* bases long (*i.e.*, there is another handle just out of sight of what is shown), will not always attach themselves to the 16-base rope needles. My calculation of the mean length of each jump rope is therefore almost certainly biased to the low side. However, *qualitatively*, there is still definitely something weird going on with the G s and even more so the C s. I don't know if that has any deeper significance in terms of the general makeup of the shadow genome, but it certainly is curious.

The other obvious ramification of the relative shortness of these runs is that I didn't actually analyze here what I set out to analyze, namely, the large repeating sections of the genome. It's possible that some of the other "Top 40" repeating sequences would get me closer to that goal, or maybe it's just something that I will never really be able to access using this needle-based method.

In any case, this diversion to the other extreme of the frequency table has given me some unexpected insights.

3. Back to growing sequences

From this point in time I will return to the sequence-growing algorithm that I described in Secs. 11 and 12 of the paper. Due to our upcoming vacation it will likely be November before I have any useful results to report to you in that direction. But I have already had some thoughts about it, which I will share here.

I originally thought that it made most sense to seek a "backbone" seed sequence that was most likely to appear exactly once, unmutated, on each of the Edith and the Vic. But having played with the data in Sec. 12 of the paper, and beyond that (with candidate unmutated seed sequences of over 110 bases, as described briefly in Sec. 3 of PR2, before being cut short by the assassination), I think that that approach is not the right one. There simply is too much data

to go through it all looking for the best unmutated seed sequence candidate, every time we want to grow another sequence.

I think it makes more sense to let all the parallel threads of the program attack the pool of available needles like a flock of seagulls attacking a discarded and unwrapped parcel of fish and chips lying on the beach. Like the heated flapping sqwarks of contention from two or more gulls who happen to grab the same chip or piece of fish, the program will need to deal with the contention of two or more threads deciding to use up the same needle: as with the gulls, one of them will need to win (well, there's no possibility of ripping a needle in two here). Apart from that issue, each thread should go about its business, trying to build the best sequence it can, dealing with both mutation splits *and* needles that appear elsewhere on the genome. Once the pile of needles has been consumed by the threads, it might be possible to then stitch together some of the resulting sequences into longer sequences. (My vision of the seagull analogy fades here, as even my best imagination can't visualize the seagulls' guts being stitched together in any way that makes the analogy worth pursuing. The disgusting sight definitely makes me want to throw away *my own* fish and chips, at any rate.)

Anyway, as I said, this is all much more delayed than I originally envisaged. My apologies for that. Maybe some of you will come up with even better suggestions by the time I get around to coding it up next month.