Summary

Overall, this seems like a valid and useful tool. However, given the presentation I'm far from 100% sure what I think they are doing is actually what they are doing. Presenting more of the raw data, such as tables comparing the consensus optimal codons identified for each AA using their method and the CEM method might help me better understand what they are doing. This gets to a larger issue of data presentation. I was especially confused by the Venn diagrams and the numbers in the overlap regions. I am also concerned as to why the looked for k-mer matches in the 3' to 5' orientation of the RNAseq data (

Basic Reporting

1. The language is overall good. (Infinitely better than my italian.) Nevertheless, there are a few errors such as
   (a) 'operative system' when the authors mean 'operating system'. This is both in the main manuscript and supplement 1.
   (b) 'list' instead of 'least' on line 168.
   (c) Not sure what the authors mean by 'elaboration time' on line 183.
   (d) Do they mean 'translation' on line 114, not transcription?
2. Run time info is sparse and at the end. Please present run time info earlier and a bit more detail, such as parallelization, if any. Through most of the manuscript I was left wondering about this. Further, there is no Supplemental File 2 with run time info.
3. I found the Venn diagrams very confusing.
   (a) I have not seen these used in this context and I suspect they are not the best way to present the data given that most of the overlap sets in the diagrams have a score of 0.
   (b) Why do the number of observations vary between the different organisms.
4. The issues with the Venn diagrams led to me concluding they've not explained things well enough for this to be published.
5. Figures text is very sparse and not informative enough.
6. The rationale for numerous steps is unclear to me. Why compare the usage of a codon between two low and high sets? This will only tell you if the usage changed. When folks talk about 'optimal' codons, its usually determined by the frequency of a codon's usage within a set of synonyms. I don't see how you determine this by simply comparing high/low gene expression usage on a single codon basis.
7. One often overlooked fact in the world of 'optimal codons' is the fact that codon usage, even in high expression genes, can be strongly influenced by mutation bias. In our work (Shah and Gilchrist 2011 PNAS, Gilchrist et al 2015 GBE) we've seen the second most efficient codon being misidentified as the 'optimal' due to the fact that the selective advantage of the 'optimal' relative to the second most efficient is small and the mutation bias favoring the second most efficient is large. This is not well appreciated and will likely only affect a few cases, but it's near and dear to my heart so I bring it to your attention.
8. Can they cite any papers that actually show using all 'optimal' codons actually leads to maximal protein expression? I know changing the usage can improve things, but the work I recall seeing where folks only use the optimal codon results unexpectedly low expression.

Experimental Design

1. Candida uses a non-standard genetic code, but since this goes unmentioned I have no idea if the authors realize this and accounted for it.

2. No real explanation is given about the multifasta protein training set. How big is it? How was it compiled.

3. Why were the RNASeq sequences searched in the reverse direction? As a biologist, that doesn't make any sense.

4. Serine is split between two separate sets in the genetic code. Did you treat these as two separate amino acids? I'm still quite confused over the numbers in the Venn diagrams and how to reconcile those with the fact that there are 18 (or 19 if you split serine) amino acids that use more than one codon.

Minor Criticisms

1. Instructions for installing biopython on linux missing. Might want to include, though most linux users are likely savvy enough to not need.

2. 'Replicate specific' is a poor choice of words.

3. When using a term for the first time in a section, please define it again.

4. I don't like the use of number of bases, I believe this is really their sample size (or related to that). As a reader, I confused it with the number of bases in the k-mer a number of times.

5. the size of a fastq file is less informative than the number of nucleotides in that file.

6. There are standard terms and symbols for kilobases, megabases which should be used in the 'The effect of the provided number of reads' (which itself is a bit awkwardly stated.)

7. You should really cite Chan and Lowe (2009) tRNA database paper instead of just the webpage.

8. Similarly you should include Peden's thesis in the references.

9. If you can run out of samples under some conditions, why not simply sample with replacement?