**Question list**

1. In <https://github.com/PapenfussLab/reads_to_domains> Github repository, the original DNA sequence file is found called “*all\_var\_exon1\_dna\_domains.fasta*”, containing 2251 sequences. Could you please show us more information about this file? For instance

(1) Are these DNA all from Rask’s paper? After checking each sequence’s identifier, there are only a few sequences which are not in Rask’s paper original dataset (gene names are “AAC47438”, 'ITvar66', 'ITvar68', 'ITvar32').

(2) Could you please explain more about each sequence’s identifier, for example “CIDRa6\_D3\_DD2var28”? Our understanding is that it is CIDRa6 domain sequence in gene DD2var28, D3 means this is the third domain. In other word, identifier is “Domain type \_ Domain order \_ Gene name”. Therefore, this file is composed of different DNA domain sequences for each gene. Is this correct?

2. I translated *all\_var\_exon1\_dna\_domains fasta* to protein sequences, and found 163 unique domains (each domain has at least three sequences)

(1) After translating to proteins, why is frame\_1 protein for every DNA tag chosen? My understanding is that you choose the frame with smallest number of “\*” as best frame by sixFrameTranslation function in mungo python module, “\*” refers to stop codon. If this best frame has zero “\*”, then make this sequence as the final translation. Is this right?

(2) It should have 163 multiple sequence alignment files. But based on the MSA files you sent us by email, there are only 150. Did you do some filtering?

3. What is your alignment tool to produce these 150 alignments from protein sequences?

A summary of our project

We aim to classify each DBLa tag to different groups (UPSA, UPSB and UPSC) in a probabilistic approach.

I notice that in your pipeline, you allocate each read to protein domains by constructing different domain’s profiles followed by searching each query sequence against these profiles to get its final domain assignment with the smallest E value. After that, in Figure 2 at Thomas Rask’s paper “Plasmodium falciparum Erythrocyte Membrane Protein 1 Diversity in Seven Genomes – Divide and Conquer”, different groups were added to each domain. Consequently, domain is the bridge between DBLa tag and UPS group. Besides, some domains in DBLa0 even also have two groups (UPSB and C) as shown in this figure, indicating domain is not conclusive to distinguish group B and C.

Therefore, we are thinking whether there is a more direct method to measure the relationship between DBLa tags and UPS groups. Currently we are seeking ways to match one query sequence to each reference sequence separately, instead of match one query sequence to a domain profile.

My own idea is to construct profiles for group A, B and C separately, like the way you construct profiles for all domains, then continue to use hmmsearch to query each tag sequence to these three group profiles and retrieve related E values. Using Rask’s data would be a good starting point.

But after trying this idea, using Rask’s data to validate its accuracy, it is only 69.97%

:(