Comments on the written PhD thesis submitted by Safonova Yana Yuryevna entitled: APPLICATION OF GRAPH MODELS TO BIOINFORMATICS ANALYSIS OF HIGHLY VARIABLE BIOLOGICAL SEQUENCES.

This PhD thesis presents a series of bioinformatics research advances and developments that are meant to improve assembly of highly polymorphic species such as the diploid and extremely polymorphic fungal species Schizophyllum Commune. Ms. Safonova and her colleagues developed a derivative algorithms dipSPAdes that masks the abundant nucleotide variants allowing better assembly and haplo-specific definition of the genome. In addition, Ms. Safonova approached the description of antibody and T-cell receptor genetic diversity in building immune products using NGS sequences to predict the sequence and structure. IN the process Ms. Safonova developed software algorithms to accomplish these purposes; 1. dipSPAdes, IgRepertoireConstructor and IgSimulator. These programs have produced important published descriptions (reference 1-4) and were applied to real data to demonstrate their utility and limitations.

The PhD thesis is written carefully and clearly to a limited extent. I say limited because there are several technical areas developed which are advanced research areas with which I am not intimately familiar, notably assembly algorithm development, particularly the intricacies of OLC and De Bruijn graph applications, immunoglobulin, antibody and T cell receptor genomic structure and organization. As such I feel that my own expertise and experience in these areas are inadequate for a detailed critique or evaluation of the technical detail that is presented here. That said, I note that the descriptions of the approaches and results have been submitted reviewed and published in high quality bioinformatics journals, leading me to suppose that the technical details have been adequately reviewed and approved not only by the mentor and colleagues but by peer reviewers of these four papers as well.

This means that my own experience in several of these areas is not so strong so my suggestions should be interpreted in that light. It would seem appropriate that experts in computational genomics, genome assembly challenges, immunology and computational immuno-genetics should be consulted for their impression as well.

I list below some impressions and suggestions that should be considered in a potential revision

- Page 4 line 15 .. …. variations, sequencing errors, computational mistakes and other artifacts….
- Page 8 last sentence in first paragraph needs citations and maybe explanation how NGS helps forensics or energy etc.
- Page 9line 6 mid 1990s I think this was closer to the ~2005
- Page 12 OLC spell out what this means
- Page 16 include maybe statistics of assembly that high accuracy you claim
- Page 17 MS data define this someplace
- Page 18 Intro paragraph makes no sense to me whatsoever.
- Page 18 2nd para line 3 what about Long indels or CNVs?
- Page 18 2nd para line 9 Human is 0.1% not 0.01 % P.
- Page 19 1st three lines is a very weird sentence Not sure I agree.
- Page 19 4 lines from bottom. How does the Bruijn graph actually mask the SNPs?
- Page 23 Why does the *Schizophyllum commune species have 10-50x* more variation than other species?
    - Is it a admixed hybrid species
    - Is repair slower?
    - Is mutation rate higher
    - Less selection? OK to review this and maybe speculate
- Page 19 4 lines from bottom. How does the Bruijn graph actually mask the SNPs?

- Page 41 Have you tried the dipSPAdes on a vertebrate like a panda or leopard which have lots of variation.?
- References I am not so fond of this format. I like to see all the authors and the authors are usually first in the reference list.
- The immunology chapter- 3 is very technical and for me difficult to follow or evaluate.

In conclusion I am happy to recommend award of Ph.D. upon revision of the thesis with attention to reviewers' suggestions.

Stephen J OBrien